Michael Kiflezghi

Full Length ERIC Sequence Occurrence in Enterobacteriophage

I. Introduction

Viruses are the most prolific biological particle on Earth [9]. It is important to characterize the genomes of these organisms to better understand how they interact with the environment and other organisms including humans. Enterobacteria such as *Escherichia coli* can play an important role in the human gut microbiome [2-4] or be pathogenic [1] depending on the strain. A previously described element of these organisms' genome is the Enterobacterial Repetitive Intergenic Consensus (ERIC) sequence. ERIC sequences are imperfect palindromes that contain several inverted repeats [6,7]. Enterobacteriophage are viruses that infect Enterobacteria. These viruses are able to acquire portions of their hosts' genomes []. The aim of this study is to determine whether or not ERIC sequences occur in the genomes of Enterobacteriophage.

II. Methods

In order to address the question of whether ERIC sequences occur in Enterobacteriophage a tool was developed using BioBIKE [11].

II.A. <u>Algorithm Development</u>

The tool, henceforth called the Eric-finder, would need to be able to identify an ERIC sequence. Several different sources have different defining characteristics of an ERIC. [6] described the defining characteristic as a set of inverted repeats at the core of the sequence (See Figure 1).

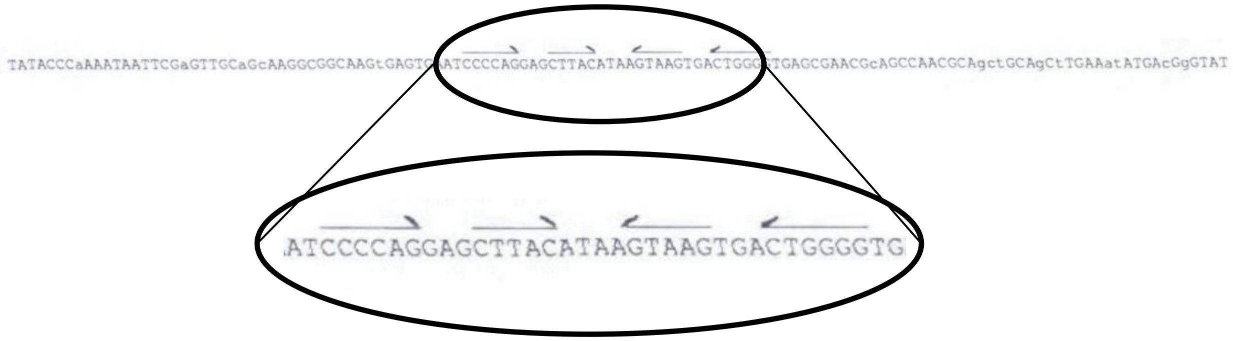
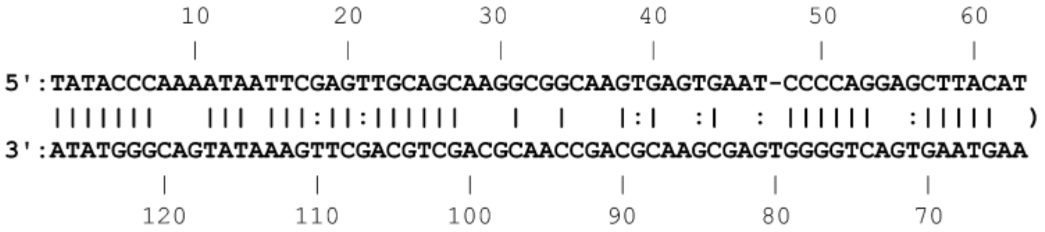

TATACCCaAAATAATTCGaGTTGCaGcAAGGCGGCAAGtGAGTGAATCCCCAGGAGCTTACATAAGTAAGTGACTGGGGTGAGCGAACGcAGCCAACGCAgctGCAgCtTGAAatATGAcGgGTAT
ATCCCCAGGAGCTTACATAAGTAAGTGACTGGGGTG

Figure 1: ERIC sequence from figure 1 of [6] illustrating position of core inverted repeat within a 127 nt long ERIC sequence

[7] characterizes ERICs as containing long terminal inverted repeats (TIRs). This is an inverted repeat occurring at the end of the ERIC sequence. [5] has a consensus sequence that contains the core inverted repeats described by [6] but also contains the TIR described by [7] (see Figure 2).



```
           10        20        30        40         50        60
            |         |         |         |          |         |
5':TATACCCAAAATAATTCGAGTTGCAGCAAGGCGGCAAGTGAGTGAAT-CCCCAGGAGCTTACAT
   |||||||   ||| |||:||:||||||   |  |   |:|  :|  : ||||||  :|||||  )
3':ATATGGGCAGTATAAAGTTCGACGTCGACGCAACCGACGCAAGCGAGTGGGGTCAGTGAATGAA
          |         |         |         |         |         |
         120       110       100       90        80        70
```

Figure 2: ERIC sequence from Wilson and Sharp. Note the terminal inverted repeat as well as the core inverted repeats beginning at positions 48 and 57.

This structural feature of core and terminal inverted repeats (see highlighted regions in figure 3) became the basis for the algorithm. In order to identify ERICs the algorithm would need to be able to identify this structure.

The Wilson and Sharp paper lists the coordinates of the ERICs found in the genome of *Escherichia coli* k-12. Only full length 127 nucleotide ERICs with perfect inverted repeats were considered from this list (see supplemental figure 1 for the list). In order to be an effective ERIC finder the algorithm must be able to first detect known instances of ERICs.

II.B. <u>The Algorithm</u>

The algorithm first instantiates a window that slides over the genome one nucleotide at a time. The window is 127 nucleotides in length. The positions of the inverted repeats within 127 nucleotide long ERICs were determined from [5-7]. The program compares the regions between positions 48-53 and 74-79. Here these regions are referred to as the outer core repeat. The next regions compared are at positions 57-61 and 66-70 followed by the terminal inverted repeats from positions 1-7 and 121-127.

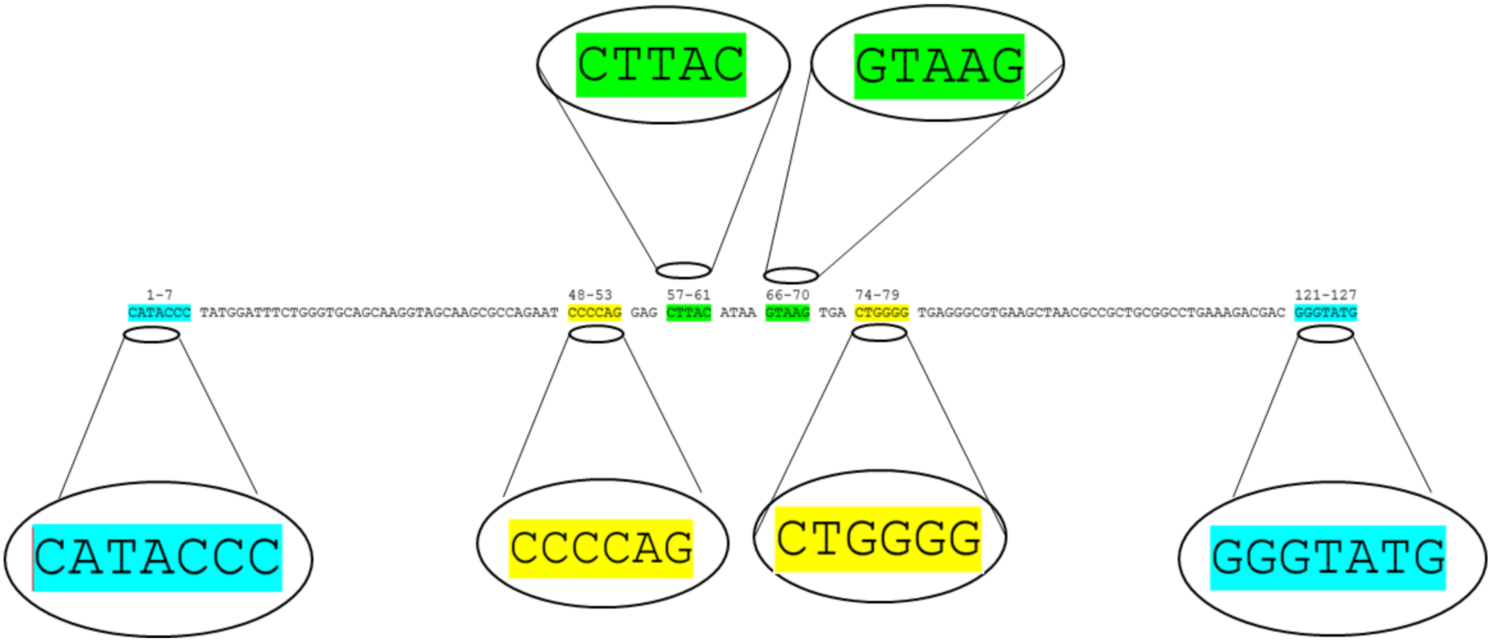


Figure 3: an ERIC sequence from the Wilson and Sharp paper. Shown is one of the sequences that contained perfect inverted repeats. See supplemental figure 1 for a full list of sequences.

A score is assigned to each region based on its reverse complimentarity (see figure 4). Each region is compared to the reverse of it's corresponding sequence nucleotide by nucleotide. When a compliment is detected, the score is incremented by 1. In early testing this approach seemed promising but allowed for many extraneous sequences when tested on the genome of *E. coli* k-12. To reduce or

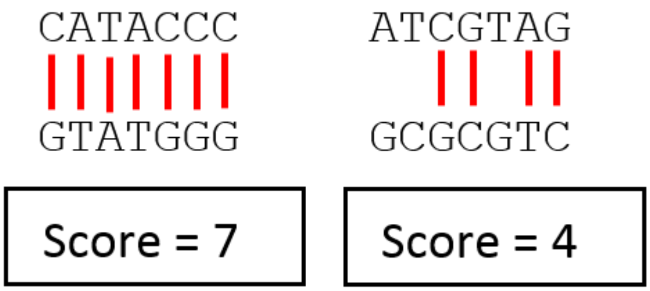


Figure 4: region scoring based on complimentarity

eliminate the extraneous results, a strict threshold was set for each region. This means that a candidate ERIC was only kept if each inverted repeat was perfectly complimentary. The program returns the sequence as well as the sequence's score. If there are no sequences within the genome being scanned that satisfy the conditions of having all perfectly complementary inverted repeats, the program returns a result of nil for that genome. The original goal was to use this ERIC-finder on all bacteria infecting viruses in the NCBI database. Given the time constraints and computational obstacles encountered, I decided to look at all Enterobacteriophage in the NCBI database. This decision was based solely on the fact that Enterobacteriophage infect Enterobacteria. I reasoned that these phage would be the most likely place for ERIC sequences to end up if they exist in any phages.

Additionally, a pattern matching tool in BioBIKE named matches-of-pattern was used to search for 100% matches of the 15 127 nucleotide long ERIC sequences described in [5].

III. Results

A total of 126 Enterobacteriophage were analyzed with the ERIC-finder and matches-of-pattern (see supplementary figure 2 for list of accession numbers).The ERIC-finder found no hits in any of the Enterobacteriophage in the NCBI database. The matches-of-pattern search found no full matches in any of the Enterobacteriophage.

These results do not mean that there are absolutely no ERIC sequences within these genomes. The strictness of the algorithm may have filtered out valid hits. The matches-of-pattern search confirms that the less complimentary ERIC sequences listed in [5] are not present in any of these genomes.

To take the search further an ERIC finder capable of detecting a wider range of ERICs without returning the extraneous sequences would need to be developed. Also, building the finder such that it can detect ERICs of varying lengths may lead to more definitive results.

References

1. Riley, L., & Remis, R. (1983). Hemorrhagic colitis associated with a rare Escherichia coli serotype. *... England Journal of ...*. Retrieved from http://www.nejm.org/doi/pdf/10.1056/NEJM198303243081203

2. Hudault, S., Guignot, J., & Servin, a L. (2001). Escherichia coli strains colonising the gastrointestinal tract protect germfree mice against Salmonella typhimurium infection. *Gut*, *49*(1), 47–55. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1728375&tool=pmcentrez&rendertype=abstract

3. Bentley, R., & Meganathan, R. (1982). Biosynthesis of vitamin K (menaquinone) in bacteria. *Microbiological Reviews*, *46*(3), 241–80. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2665567

4. Reid, G., Howard, J., & Gan, B. S. (2001). Can bacterial interference prevent infection? *Trends in Microbiology*, *9*(9), 424–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22901538

5. 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

6. Hulton, C. S., Higgins, C. F., & Sharp, P. M. (1991). ERIC sequences: a novel family of repetitive elements in the genomes of Escherichia coli, Salmonella typhimurium and other enterobacteria. *Molecular Microbiology*, *5*(4), 825–34. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1713281

7. Gregorio, E. De, Silvestro, G., Petrillo, M., Carlomagno, S., Paolo, P., Nocera, D., … Nocera, D. (2005). Enterobacterial Repetitive Intergenic Consensus Sequence Repeats in Yersiniae : Genomic Organization and Functional Properties Enterobacterial Repetitive Intergenic Consensus Sequence Repeats in Yersiniae : Genomic Organization and Functional Properties. doi:10.1128/JB.187.23.7945

8. Weisberg, R. a, & Adhya, S. (1977). Illegitimate recombination in bacteria and bacteriophage. *Annual Review of Genetics*, *11*, 451–73. doi:10.1146/annurev.ge.11.120177.002315

9. Edwards, R., & Rohwer, F. (2005). Viral metagenomics. *Nature Reviews Microbiology*, *3*(June), 801–805. Retrieved from http://www.nature.com/nrmicro/journal/v3/n6/abs/nrmicro1163.html

10. Morris, P., Marinelli, L. J., Jacobs-Sera, D., Hendrix, R. W., & Hatfull, G. F. (2008). Genomic characterization of mycobacteriophage Giles: evidence for phage acquisition of host DNA by illegitimate recombination. *Journal of Bacteriology*, *190*(6), 2172–82. doi:10.1128/JB.01657-07

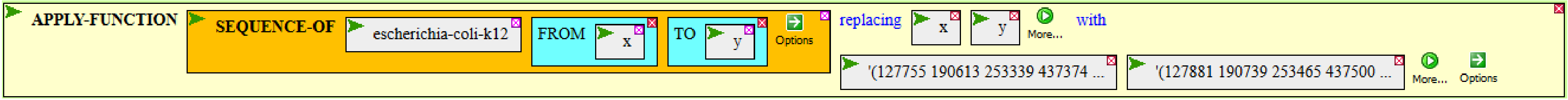
11. Elhai, J., Taton, A., Massar, J. P., Myers, J. K., Travers, M., Casey, J., … Shrager, J. (2009). BioBIKE: a Web-based, programmable, integrated biological knowledge base. *Nucleic Acids Research*, *37*(Web Server issue), W28–32. doi:10.1093/nar/gkp354

Michael Kiflezghi

Supplemental Materials

1-7		48-53		57-61		66-70		74-79		121-127
CATACCC	TATGGATTTCTGGGTGCAGCAAGGTAGCAAGCGCCAGAAT	CCCCAG	GAG	CTTAC	ATAA	GTAAG	TGA	CTGGGG	TGAGGGCGTGAAGCTAACGCCGCTGCGGCCTGAAAGACGAC	GGGTATG
AATTTCC	TTCGTCTTTCACGCCATAGCGGCGTTGGCGTCGCCCGCTC	ACCCCG	GTC	ACTTA	CTTG	TGTAA	GCT	CCCGGG	GATTCACAGGCTAGCCGCCTTGCTCTGACGCGAAATACTTC	GGAAATT
CTCCCCC	AAAATAGTTCGAGTTGCAGAAAGGCGGCAAGCTCGAGAAT	TCCCGG	GAG	CTTAC	ATCA	GTAAG	TGA	CCGGGA	TGAGCGAGCGAAGATAACGCATCTGCGGCGCGAAATATGAA	GGGGGAG
TATACTC	TAAATAATTCGAGTTGCAGGAAGGCGACAAGCGAGTGAAT	CGCCAG	GAG	CTTAC	ATAA	GTAAG	TGA	CTGGGG	TGAACGAACGCAGTCGCAGTACATGCAACTTGAAGTATGAC	GAGTATA
TATACTC	GTCATACTTCAAGTTGCATGTGCTGCGGCTGCATTCGTTC	ACCCCA	GTC	ACTTA	CTTA	TGTAA	GCT	CCTGGG	GCTTCACTCGTTTGCCGCCTTCCTGCAACTCGAATTATTTA	GAGTCTA
TATACTC	GTCATACTTCAAGTTGCATGTGCTGCGTCTGCGTTCGCTC	ACCCCA	GTC	ACTTA	CTTA	TGTAA	GCT	CCTGGG	GATTCACTCGCTTGTCGCCTTCCTGCAACTCGAATTATTTA	GAGTATG
TATTCTC	GTCATACTTCAAGTTGCATGTGCTGCGTCTGCGTTCGCTC	ACCCCA	GTC	ACTTA	CTTA	TGTAA	GCT	CCTGGG	GATTCACTCGCTTGTCGCCTTCCTGCAACTCGAATTATTTA	GAGTATA
TATACAC	AAAATCATTCAAGTTGCATCAAGGCGGCAAGTGAGCGAAT	CCCGAT	GAG	CTTAC	TCAG	GTAAG	TGA	TTCGGG	GGAGCGAACGCAGCCAAGGCAGAGGCGGCTTGAAGGATGAA	GTGTATA
TATACAC	TTTATCCTTCACGCTGCCTCTTCGTTGACTGCCTTCGCTC	ATCCCA	TTC	ACATA	GTTA	TCTAT	GCT	CATGGG	AGTTCACTCAGTTGCCGCCTCGATGCAACGCGAATGATTTC	GTGTATT
TCCGCTA	AATGATTCGCGTTGCAGGAAGGCGGCAAGTGAGTGAAGCC	CCAGGA	GCA	TAGAT	AACT	ATGTG	ACT	GGGGTG	AACGAGCGCAGCCAACGCATCTGCGGCGTGAAGCATGACGC	GGAAATT
TACTCGT	CATACTTCAAGTTGCATGTGCTGCGTCTGCGTTCGCTCAC	CCCAGT	CAC	TTACT	TATG	TAAGC	TCC	TGGGGA	TTCACTCTCTTGTCGCCTTCCTGCAACTCGAATTATTTAGA	GTATGAA
CACCAGC	TGTTTGCCCTGTACGGCATCGAAGCGACGCTGTTCATAAC	GCGGCG	TAA	TACCG	TTTT	CTTCA	GGC	ATGATC	CAGATCTGATACAGATGCAGACGCTCGGTGCTGCTTGGGTT	GTACTCT
ATCGTAG	TTAAAGACGTGCGTCACTGCCGGAATATGCAAACCACGCG	CGGCAA	CGT	CGGTG	GCAA	CCAGA	ATA	TCCAGA	TCGCCACGGGTAAATTCATCAAGAATACGCAGACGTTTTTT	CTGCGCG
CCTGTTC	CGTATTGGTCGTGGACGTGCGCCGACTGGCGAACCTGCGG	CGGCAG	CGG	AAATG	ACCA	AATGG	TTT	AACACC	AACTATCACTACATGGTGCCGGAGTTCGTTAAAGGCCAACA	GTTCAAA
GTCTCTT	TCCATGCTTTGCGCAGGGAAGATTCCTCAAAGTGCTGGCG	GTCAAA	CCA	CTCCT	GTAG	CTCGA	CCA	GCCCTT	TACGGGTGAGATCGCGCGGGCGATTAATAACTGCCTGCAAT	GCCGGTT

Supplemental Figure 1: ERIC sequences obtained using the coordinates provided in [5] using:



APPLY-FUNCTION SEQUENCE-OF escherichia-coli-k12 FROM x TO y

replacing x y with '(127755 190613 253339 437374 ... '(127881 190739 253465 437500 ...

Note lack of complimentarity in some of the highlighted regions. These sequences are derived from a pattern matching allowing for some degree of dissimilarity to an ERIC sequence.

“NC_019485” “NC_019524” “NC_019423” “NC_023561” “NC_011045” “NC_015249” “NC_022968” “NC_000924” “NC_011040” NC_004813” “NC_001426” “NC_019500” “NC_019920” “NC_014662” “NC_021315” “NC_010583” “NC_006949” “NC_011042” “NC_004301” “NC_001420” “NC_002166” “NC_019768” “NC_019710” “NC_019717” “NC_019714” “NC_019769” “NC_019767” “NC_019724” “NC_019711” “NC_019723” “NC_019719” “NC_002167” “NC_018855” “NC_019922” “NC_001332” “NC_007856” “NC_007817” “NC_014260” “NC_019501” “NC_001954” “NC_002014” “NC_007291” “NC_019419” “NC_012741” “NC_010105” “NC_012740” “NC_008152” “NC_007637” “NC_007456” “NC_015719” “NC_019707” “NC_003287” “NC_001417” “NC_010237” “NC_000929” “NC_001901” “NC_018835” “NC_005856” “NC_001895” “NC_002371” “NC_001609” “NC_001421” “NC_009821” “NC_010324” “NC_005340” “NC_001890” “NC_012638” “NC_014467” “NC_008515” “NC_007023” “NC_005066” “NC_012635” “NC_004928” “NC_007603” “NC_004831” “NC_015269” “NC_012223” “NC_005841” “NC_003444” “NC_012868” “NC_005833” “NC_003298” “NC_000866” “NC_005859” “NC_001604” “NC_009540” “NC_020414” “NC_000902” “NC_007821” “NC_011356” “NC_001330” “NC_009514” “NC_022750” “NC_019503” “NC_001416” “NC_019706” “NC_019708” “NC_019704” “NC_019716” “NC_019709” “NC_019705” “NC_010106” “NC_003356” “NC_001422” “NC_019517” “NC_014792” “NC_019399” “NC_019403” “NC_019404” “NC_019718” “NC_019526” “NC_019720” “NC_019715” “NC_019721” “NC_018086” “NC_017732” “NC_012419” “NC_015270” “NC_023551” “NC_023595” “NC_009904” “NC_013696” “NC_013646” “NC_013643” “NC_013648” “NC_013644”

Supplemental Figure 2: List of NCBI accession numbers for all Enterobacteriophage in the NCBI database