

Introduction

Bacteria as a domain are highly resilient organisms. Struggles for food, resources and hospitable living conditions are a constant battle for bacteria, but they manage to proliferate at extreme rates despite these limiting conditions. However, one particular threat can be highly damaging to a bacterial population, even in an otherwise friendly environment. That threat is the phage. Specifically a bacterial virus, phages have incredible population size, with an estimated 10^{30} total phages in the biosphere^[1] (that is one million trillion trillion). In fact, samples of sea waters show up to 70% of observed bacteria to be infected with phages, with 900 million virions existing per milliliter of sea water^[2].

In order to deal with this constant and ubiquitous threat, bacteria have evolved a number of ways to combat phages. One particularly interesting strategy bacteria employ is one that is homologous to the B cell-mediated, antibody-based adaptive immune system found in mammals. However the adaptive immune system of the bacteria is based immediately off of the bacteria's own genome. This immune response is called a CRISPR, or a "clustered regularly interspaced short palindromic repeat". These are regions in the DNA which have repeating sequences 23-47 base pairs long, interrupted by variable regions of similar or slightly longer length^[3]. These two regions which sandwich each other are referred to as direct repeat sequences, and variable spacer sequences.

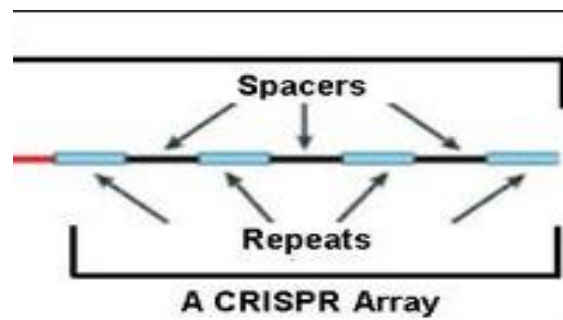


Figure 1. Diagram of a basic CRISPR on a bacterial genome. Courtesy of American Society for Microbiology website, <http://schaechter.asmblog.org/schaechter/2011/04/six-questions-about-crisprs.html>

CRISPRs function by transcribing a spacer sequence into RNA and associating it with a Cas (CRISPR-associated protein). Via this method, the CRISPR can effectively act as a search tool to find DNA sequences like itself^[4]. What it is searching for is phage DNA. If the CRISPR finds a match, it can tag the phage DNA for deletion, and potentially save the bacteria from infection.

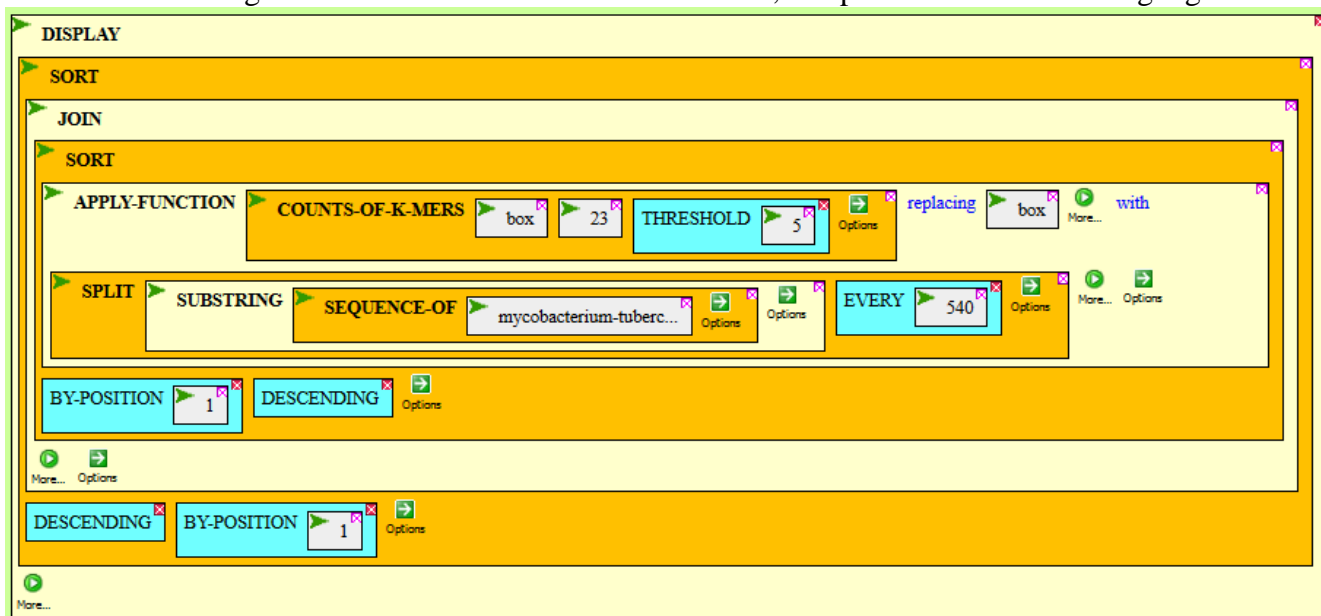
However, the bacteria needs to know what to search for first. The variable spacer sequences come directly from free DNA the bacteria has previously encountered, and assimilated into its genome. While this DNA may be from non-phage sources, it is often phage DNA^[5]. By this method, if a bacteria has encountered phage DNA and not been killed, it may gain active immunity towards that phage, as well as all of its exponentially reproducing descendants.

Due to this consideration, there should be large similarities between bacteria CRISPR variable spacer sequences and actual phage DNA.

To look into this, it was my goal to first find CRISPR sequences. My next goal was to, after finding such sequences, compare the variable spacer sequences to known phage DNA. My question was twofold: If I could find a CRISPR variable sequence that matches to the genome of a reasonable phage candidate, and second: If the collection of variable regions in CRISPRs has any bias towards certain characteristics of a virus genome (e.g. coding vs noncoding regions).

Methods

As previously mentioned, the first step is to actually find a CRISPR sequence. To do this, I randomly selected various organisms from the Biobike database. Next, I implemented the following algorithm:

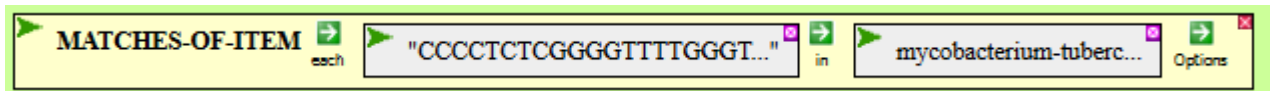


This breaks the bacterium's genome into chunks of 540. I chose this box size for two reasons. 1, previous work on a completely unrelated project happened to come across a CRISPR sequence by happenstance, and a box size of 540 was being used then (with slightly better justification as well). The second reason is that 540 is actually a fairly good search size. CRISPR sequences appear in tight clusters in the genome, not spread throughout the chromosome. Assuming a maximum sized CRISPR, with a direct repeat sequence of 47bps and a variable spacer sequence of 60 bps, I should still return at least five results.

The program then, in each 540 chunk, executes the “counts-of-k-mers” function. This tallies up nucleotide strings of 23bps in length that repeat multiple times. Here, I set 23 as the search size as this is the lower limit for CRISPR direct repeat regions. I also set the threshold to 5, as this would return more likely results (as previously stated, 5 should be the minimum return even for a large CRISPR). The compiled list was then joined and sorted to be more easily viewable. In this example, using *Mycobacterium-tuberculosis-H37Ra*, a maximum result of 7 was obtained.

COUNT	WORD
7	CCCCCTCGGGGTTTTGGGCTCG
7	CCCTCTCGGGGTTTTGGGCTCGA
7	CCGTCCCCTCTCGGGGTTTTGGG
7	CCTCTCGGGGTTTTGGGCTCGAC
7	CGTCCCCTCTCGGGGTTTTGGGT
7	CTCGGGGTTTTGGGCTGACGAC
7	CTCTCGGGGTTTTGGGCTGACG
7	GTCCCCCTCTCGGGGTTTTGGGTC
7	GTTTCCGTCCCCTCTCGGGGTTT
7	TCCCCTCTCGGGGTTTTGGGCT
7	TCCGTCCCCTCTCGGGGTTTTGG
7	TCTCGGGGTTTTGGGCTGACGA
7	TTCGTCCCCTCTCGGGGTTTTG
7	TTTCCGTCCCCTCTCGGGGTTTT
COUNT	WORD
6	CCCCCTCGGGGTTTTGGGCTCG
6	CCCTCTCGGGGTTTTGGGCTCGA
6	CCGTCCCCTCTCGGGGTTTTGGG
6	CCTCTCGGGGTTTTGGGCTCGAC
6	CGTCCCCTCTCGGGGTTTTGGGT

From here, the sequence can be selected and searched for in the bacterium's genome using matches-of-item.



The specific instance above (highlighted) contained a highly probable CRISPR. The following is the returned sequence from the bacterium. It is also interrupted by an unknown protein. This is interesting, but has little to do with the topic at hand.

ATGCTCATCCTGAATGCCGGTCAACAGACGCGGTGGCGACCCAGTCGTCGTA **STTTCCGTCCCCTCTCGGGGTTTTGGGCTCG**
ACGACTCGGGCACGGCCGAAACACCGCGCAAGGGCGGTTCAAGTTTTCCGTCCCCTCTCGTGGTTTTGGGCTGACGACTGGG
AGGATGTCACTCGGACATAGCTGTGCATCGGCGGTGTGTTTTCCGTCCCCTCTCGGGGTTTTGGGCTGAGGACATGGAGCAGTA
GCGTGGCTGTGGTGTGGCGGGCGATATGC **STTTCCGTCCCCTCTCGGGGTTTTGGGCTGACGAC**TGCTGCACCTCCCGCACC
CGGTGCGATTTCTGCGTCCAGTTTTCCGTCCCCTCTCGGGGTTTTGGGTCCGACGACCCCGATAGTCGCGCTCGTCCATGTCCCA
CCATGAGG **STTTCCGTCCCCTCTCGGGGTTTTGGGCTGACGAC**TACCTGATAGAAGCCGAAAGCTCCGTGCCGTGAG **STTT**
CCGTCCCCTCTCGGGGTTTTGGGCTGACGACAGGGCACTGGACCTGTATGAGGCACAGATGGCGTACTA **STTTCCGTCCCCT**
CTCGGGGTTTTGGGCTGACGACCCGGATCGGTTACCCACGCCGATTTACTGGCCATCGTCGG **STTTCCGTCCCCTCTCGGG**
TTTTGGGCTGACGACACTTGCGCACAAACGCATCCGCCATCCAGGGGC **STTTCCGTCCCCTCTCGGGGTTTTGGGCTCGA**
CGACCTGAAAGGGGACTGTGGACGAGTTCGCGCTCAAAAT **STTTCCGTCCCCTCTCGGGGTTTTGGGCTGACGAC**TTGAAC
ACGCCGATACTATTTGGTTCGGGAGTGATAAA **STTTCCGTCCCCTCTCGGGGTTTTGGGCTGACGAC**CGGACTTGATCGACG
CGAACCTGTCTGACGCGAACCT **STTTCCGTCCCCTCTCGGGGTTTTGGGCTGACGAC**GGCTGGAAAAAGGGCGCGGGGCAACC
GCATCGTCAAGA **STTTCCGTCCCCTCTCGGGGTTTTGGGCTGACGAC**GCGTTGTGGTCTGTGCTGAGGACCTGTATTTTCGCT
G **STTTCCGTCCCCTCTCGGGGTTTTGGGCTGACGAC**CATAGTGTGGTGTGTGATCGCTAAACGCCGGGGCA **STTTCCGTCC**
CCTCTCGGGGTTTTGGGCTGACGACCTATCCGCGGAAGAGATCACGAATCCGGCGTCGAAGG **STTTCCGTCCCCTCTCGGG**
STTTTGGGCTGACGACATGCTGAGCTGAGGCGCCGATGATGGTGGTGTGTAAG **STTTCCGTCCCCTCTCGGGGTTTTGGGT**
CTGACGACTGACAGGGTGCGGTGGTCTGATCGGCTCCCAGATTTCCGTCCCCTCTCGGGGTGAACCGCCCCGGTGAGTCC
GGAGACTCTCTGATCTGAGACCTCAGCCGGCGGCTGGTCTCTGGCGTTGAGCGTAGTAGGCAGCCTCGAGTTCGACCGGCGGG
ACGTCCGCCAGTACTGGTAGAGGCGCGGATGGTTGAACCACTCGACCCAGCGCGGGTGGCCAACTCGACATCCTCGATGGA
CCGCCAGGGCTTGCCGGGTTTTGATCAGCTCGGTCTTGTATAGGCCGTTGATCGTCTCGGCTAGTGCAATTGTCATAGGAGCTTC
CGACCGCTCCGACCGACGGTTGGATGCCTGCCTCGGCGAGCCGCTCGCTGAACCGGATCGATGTGTAAGTACTGAGATCCCCTATCC
GTATGGTGGATAACGTTCTTTCAGGTCGAGTACGCCCTTCTTGTGGCGGGTCCAGATGGCTTGTCTGATCGCGTTCGAGGACCAT
GGAGGTGGCCATCGTGGAAAGCGACCCGCCAGCCAGGATCCTGCGAGCGTAGGCGTCCGGTGACAAAAGGCCACGTAGGCGAACC
CTGCCAGGTGACACATAGGTGAGGTCTGCTACCCACAGCCGGTTAGGTGCTGGTGGTCCGAAGCGGCGCTGGACGAGATCG
GCGGGACGGGCTGTGGCCGGATCAGCGATCGTGGTCTGCGGGCTTTGCCGCGGGTGGTCCCGGACAGGCGGAGTTTGGTCA
CAGCCGTTGACGGTGCATCTGGCCACCTCGATGCCCTCACGGTTCAGGTTAGCCACACTTTGCGGGCACCGTAAACACCGT
AGTTGGCGGCGTGGACGCGGCTGATGTCTCCTTGAGTTCGCCATCGCGCAGCTCGCGGCGGCTGGGCTCCCGGTTGATGTGG
TCGTAGTAGGTGATGGGGCGATCGGCACACCCAGCTCGGTCAGCTGTGTGCGAGATCGACTCGACACCCACCGCAAACCATC
GGGGCCCTCGCGGTGGCCCTGATGATCGGCGATGAACCGGGTAA **TTAGCGTGTGGCCGGTTCGAGCTCGGCCGCGAAGAAAGC**
CGACGCGGTCTTTAAAAATCGCGTTCGCCCTTCGCAATTCGGCGTGTGCCGCCGAAGCGCTTCAGCTCAGCGGATTCCTTCGG

TCGTGGTCCCGGGCCGTGCGCCGGCATCGACCTGCGCCTGGCGCACCCACTTACGCACCGTCTCCGCGCAGCCAAACACCAAGT
 AGACGGGCGACCTCACTGATCGCTGCCCACTCCGAATCGTGCTGACCGCGGATCTCTGCGACCATCCGCACCGCCCGCTCACG
 CAGCTCCGGCGGGTACCTCCTCGATGAACCACCTGACATGACCCCATCCTTTCCAAGAACTGGAGTCTCCGGACATGCCGGGG
 CGGTTACAGGTTTTTGGGTCTGACGACTCGCGGCGAGCACGTCTCACCCAGCAGGCGGTGAGGTTGGGTTCCGTTCCCTCTCG
 GGTCTGACGACTTGTCTCAATCGTGCCGTCTGCGGTGACACGTCCAACTTTCCGTTCCCTCTCGGGGTTTTTGGGTCTGACGA
 CACCAGGATCAGCCGAAGCCAGTTAGCGCAATCCAAGTTCCGTTCCCTCTCGGGGTTTTTGGGTCTGACGACTCCCGGACC
 ATCTGCAGCTCGCCGGTCCATGCGTTCCGTTCCCTCTCGGGGTTTTTGGGTCTGACGACTCGGAGTCATCCGCGCGGGCCG
 GCGGATTTGTTGCCGGTTCCGTTCCCTCTCGGGGTTTTTGGGTCTGACGACTGGCGATTTACGACGCTGACGGGAACCTCGTG
 CGAATGTTTCCGTTCCCTCTCGGGGTTTTTGGGTCTGATCCGCGAAATTTACTGCGCGTTATTCAAGTTCCGTTCCCTCTCG
 GGTCTGACGACTCCGAGCCGACCATCCGCATCACACCGAAAGGGTTGGCGCAAAGTTCCGTTCCCTCTCGGGGTTTT
 TGGGTCTGACGACTACGTGGGGAGAGGGAATGGCAATGATGGTCGACGAAAGTTCCGTTCCCTCTCGGGGTTTTTGGGTCTGACG
 ACTCGGACAGCATCTCCCGGGCGGGCAGCAGATATCCCATTTCCGTTCCCTCTCGGGGTTTTTGGGTCTGACGACTCGACC
 CGTGGCCGCCAGGTTGCCGCCCGCTTGTCTACCTGTTCCGTTCCCTCTCGGGGTTTTTGGGTCTGACGACTCGGAAAGTCAA
 CTAGAGCGGGTGTGAAACGCTGCCGTTCCGTTCCCTCTCGGGGTTTTTGGGTCTGACGACTATGCGAATCCGCTGTGACGAC
 ATGGGATCCGAGTTCCGTTCCCTCTCGGGGTTTTTGGGTCTGACGACTTAGCGGCCCGCGGAGGCTGGGGCGGTTTC
 ACGCTCCGTTCCCTCTCGGGGTTTTTGGGTCTGACGACTAGGCTGAAATTTGAAGCCGGAAATGACGACGCAATTTGGTGT
 TCCGTTCCCTCTCGGGGTTTTTGGGTCTGACGACTTAAAGCCCCGCTTAAATCCCCGCACAAAAGTTGGGTGAGAAAAAGGTT
 TCCGTTCCCTCTCGGGGTTTTTGGGTCTGACGACTGATGATTTGGTGGCGTATGACGTGCTACTGAGGTGTTTCCGTTCCCT
 CTGAGGTTTTTGGGTCTGACGACTTAGAAGGCGATCACTGGAAGCACGGCGCTTGCGAAGTTCCGTTCCCTCTCGGGGTTTT
 TGGGTCTGACGACTTGGTCAAAAAGCTGTGCCCCAAGCATGAGGCAAAAAAGTTCCGTTCCCTCTCGGGGTTTTTGGGTCTGACG
 ACTAGGAGGAGCGTGTATCCAGAGCCGGCGACCCCTATGTTCCGTTCCCTCTCGGGGTTTTTGGGTCTGACGACTGTGCAA
 GAAATCCGGGTTGCAAGTGAACACGGTTTTTAAAGTTCCGTTCCCTCTCGGGGTTTTTGGGTCTGACGACTCTATGGACAATTCG
 TACAGCGTGTGGTAAACAATGCTGTGATGATGTCAAAAGAACAACAACTCCTCTGCGCTGACAAGCCGTTCCCTCTCCGTTAG
 ACGTAACTGCCGCAACACCTCTTATCTTATAGATCCGGATGTTGTGCGAGTCGATGGCGAAGCGGTGCGATACGTGCAACTAGT
 TTCGCGAGCTGGCCCTTCGTCAGCATCGCTTCGAATGCGGACTC

As can be seen, the direct repeat sequence is 36bps in length. Based on Biobike's knowledge of protein function, the sequence is also directly before a CRISPR protein.

```

3135481 TGGTCTGACGACGCTGCAAGAATCCGGGTTGCAGTGCAACACGTTTAAAGTTCCGTC
3135541 CCTCTCGGGTTTTTGGGTCTGACGACTCTATGGACAATTCGTTCCAGCGTGTGGTAACAA
3135601 TGCCCTGCTGATGATGTCAAAAGAACACAACTCCTCTGCGCTGACAAAGCCGTTCCCTTCC
3135661 GTAGAACGTAACCTGCCGAACACCTCTTATCTTATAGATCCGGATGTTGTCGCACTCGAT
3135721 GCGAAGCGGTGATACGTTGCAACTAGTTTCCGAGCTGGCCCTTCGTCAGCATCGCTTC
3135781 GAATGCGGACTCTTGGACCGGATAGCCAAACCCGGCCAGGATCTTCGCAAGTGAAGCCCG
3135841 CCGCCGTTGTCGCTGATGTCGTAATATTACGAGGACGAACATCTGCCTATAGTGGCCGT
3135901 GGACTCGTCCACTTTGAGCGGGAGATTGAAGTACTCCTCACGGCTGCGAGTGGGCATTTA
3135961 GGCTCCGGATGGCTCGGAGGTGATATCGATATCGACGAGCCGCGAGGGTGCCCGGCTTC
3136021 GATAACACGACGAGGCTTTGCAAGTTCGAGGCGTACTGAAAGTGTATCGGTG
3136081 AGGATCGCCTTTGATGATAGGTGGCGGTTGTCGATTCGATACCAAGGCGCGCGCAT
3136141 GGATCGTGGGCTTCCCGTGTGCGAAGACGGCCCGTGTGCGAGTTCTTGTGAAAGC
3136201 CCGGTTGTCGACCACCCGTTCCGCGATCAATCGAAGTACGGTGTATCGATGATCGGCGC
3136261 CCGCCATACCTCCATGAGGTGCTCGCCAAACGTTGCGTGCCTCGTGAATCCTGTTGATG
3136321 GAAACGATATACGCGTTTCCGCTGTGACGCTCGATCGCCCTATGATGTTCTGTACAG
3136381 CAGCGAATAGCCGAGGCTGACCATCGAGTTGAAGGCGTCCAAACGCGCGCGAGTCCAGCG
3136441 CCGCTGGAATGCGAATCCTCGCGGACGAGATGCCCGAGCGCGGTGAAGTATGCCCTTGC
3136501 GGCATTTCCCTCGAACCCGTTCAACTCCGCGAGGAGCCCGATCGATCGACCCAGGCCAG
3136561 CGAGTGTCTCATCGTGCAGGATGCTCTCAGCAACGTTTCCGCGAGCTGTGTGCCCGAAT
3136621 CAAGGCTGTGATTCAGGATCTTCTCGACACGATCCGCTTGTCTAACGACAGGCGAAA
3136681 CGCAGGATCGTGGTGGGTAACCTTGTGACGAGCCGCGCGCGTATGACACGTCGGG
3136741 TGTGAGATCCGCCCTGGTAGTGGCCGTCGCTCGTGAAGAGCTGGATGTCGCGCTCAG
3136801 CTTGAGCATCTAACGATGAAGGGGCTTGTATCGTCCGCGCCCAACAGCGTGTGTC
  
```

Mtub-H37Ra.Mtub-H37Ra-7901 (3135616 <- 3135957)
 CRISPR-associated protein Cas2

Mtub-H37Ra.Mtub-H37Ra-7902 (3135958 <- 3136974)
 CRISPR-associated protein Cas1 [SS]

So, this region is very likely to be a CRISPR. The method of finding CRISPRs appears to be effective.

I repeated this with other bacterial organisms.

Nitrosomonas europaea, a nitrogen fixing bacterium found in soils, returned this:

```

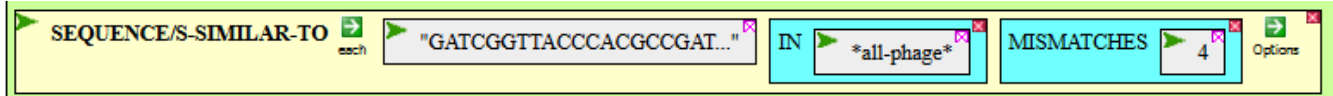
AACGCATTTACGCACTAATCCGGCGGTAATTCAGGGTATCGGTGTTTTCGGTATCCTGCTGTGAAAAAG
ATAGGTGGGCGTTGTCTCAATCCCTTTGAAATCAGGGCATCGGTGTTTCATCTATTGGCCGATAGGTTT
CGAAGTAAAGATGGTCTCAATCCCTTTGAAATCAGGGCATCGGTGTTTCCTACTGGTTCGAGACATAT
GAGGAAAAGTATTTGTCTCAATCCCTTTGAAATCAGGGCATCGGTGTTTCAAACCAGAATCAAGAAAGA
AGACAAAAATTACTACGTCCTCAATCCCTTTGAAATCAGGGCATCGGTGTTTCAAACCAGAAATCAAGAAA
GAAGACAAAAATTACTACGTCCTCAATCCCTTTGAAATCAGGGCATCGGTGTTTCCCTTGGCTAACGCGA
  
```

And C. tetani, which causes tetanus, returned this:

```
GGGAGTAGACAATTATAGTTATATTAATGGTTTATAAATTATTTAACATAACTTGTATTATAATA
ATAGACTATAAGAAATCCCAACTTAGTAGGTTAGACAGAAATAGAAGTTATACTATAGACATAT
AGTGCATCATA GATGAACGC GTATTAGTAGCACCATATTGGAATGTAAATTTAAATGAAAGGTA
CTAAATTTAAGGTAAGAATGGTGGTATAGTAGCACCATATTGGAATGTAAATAGCATTCCCTC
TATCTCCATTAECTACTGAAAAAGGAGTATTAGTAGCACCATATTGGAATGTAAATGCTATAA
GAATAATTCTAATTTATCCAAGGAACGTGGTATTAGTAGCACCATATTGGAATGTAAATTTCC
AATTTACTAGCTGCTACCCCAACGCCTAATAATGTATTAGTAGCACCATATTGGAATGTAAAT
TTAAATATATTACTTCTTCTGCACTGTAGGTTTTCGTATTAGTAGCACCATATTGGAATGT
AAATTTGTTTCGTAAGTGTAAAGCTATCTTTCTTATGCGTATTAGTAGCACCATATTGGAATG
TAAATAGAGCCTAGTTTCTAAGCCCTTATAACCAACTTACCGTATTAGTAGCACCATATTGG
AATGTAAATATACAATGCTCCATGGAAAGGACTCCACTTAGATATATA GTATTAGTAGCACC
ATTGGAAATGTAAATCCCACATCATAAAGGATATAAAAATTACCACCTTCCGTATTAGTAGCA
CCATATTGGAATGTAAATGTTTTAATATTAATATCGGCAAGTGCTAATTCATATGGTATTAGT
GCACCATATTGGAATGTAAATGTTTTAATATTAATATCGGCAAGTGCTAATTCATATGGTATT
GTAGCACCATATTGGAATGTAAATAACTGCACAGTATCACCGCTAGCTTTTAATTCCTTAGT
TTAGTAGCACCATATTGGAATGTAAATTTTTGAAGTATATTATAAAGGCACAGTAACACGCC
GTATTAGTAGCACCATATTGGAATGTAAATGCTTAACTCTTAAAAAAGATAAAGTTCTAAAT
TCGTATTAGTAGCACCATATTGGAATGTAAATAAGATGCAGCCAACGCACCTGGATATATGGC
TTTGGGTATTAGTAGCACCATATTGGAATGTAAATGGTAATGTAAATTAATTCTACAACCAATA
ATAGCAATGTATTAGTAGCACCATATTGGAATGTAAATATACAGAATACAAGATTATAGTTAGT
GGATATAGAA GTATTAGTAGCACCATATTGGAATGTAAATAATGAAGTCAAATAATAACATAC
CATTTGTGCTCGTATTAGTAGCACCATATTGGAATGTAAATTTAAATCTGGTTTATTTTTTA
CATTTCTCCAATCCGTATTAGTAGCACCATATTGGAATGTAAATAAGTGCCTTATTACGCCCTT
CTATATGTCGGAATACCGTATTAGTAGCACCATATTGGAATGTAAATTTTTATTAGCAATCC
TTTGTACTGCCATCTTCCGGTATTAGTAGCACCATATTGGAATGTAAATCCTAGTACGCCAG
CATACCCAAAAAAGAACTACTTAA GTATTAGTAGCACCATATTGGAATGTAAATCTGGAAAG
AGGCAATAAAGCATTAGGAATAATAAAATGGTATTAGTAGCACCATATTGGAATGTAAATCTT
ACTAACACTTTAGACCTAGTATTAATAAATTTTGTATTAGTAGCACCATATTGGAATGTAA
ATGTGTACATCTCCCAATTTCTCTCATAATACTTTAAGTATTAGTAGCACCATATTGGAATG
TAAATGCTATAGCTAGTAGATAGATACGTTGCGAGAATGGGTATTAGTAGCACCATATTGGA
ATGTAAATAAATTAATTGGCAGTATATGCTATACCATCTATAGCGTATTAGTAGCACCATATTG
GAATGTAAATATCTAACTCAATATTTCTTCTTTTACATCCTGTTTAGTATTAGTAGCACCATA
TTGGAATGTAAATAAATAATAAAGATAGTAGGTTAAAGGGTATATTAGTATTAGTAGCACC
ATATTGGAATGTAAATTTATAAAATTTCTATTTCTAGTTCTTCTTGAGTATATTATTAGTAGC
CATATTGGAATGTAAATGTCGACCCATTGGAGTTAGACAGATGGGATTTTCAGTATTAGTAGC
ACCATATTGGAATGTAAATTTAGCATCTATAATATTACTTCTTTAATAGTTCCTGTATTAGTAG
CACCATATTGGAATGTAAATATATGTGATGATGAATTAGAGAAAGTCTTGAAAGGTATTAGT
AGCACCATATTGGAATATAAATGACTGTAACCTATAGTATAAATAAGTAAATAGGGACGGTTA
ATGAATTTTTCAAAAGTATGATATAATAAAAACATAATTTACTTGATGAGGTATGATATGGATT
TTAATGAAACAATATATAATAGAATATTAATAATTTGTTAAGGAGAAATCCCGATAGTGTATT
GCCACAGGATTTTGTAGAACGGGGAAAGAAAGATGCATATTTATTTTAATGCGAAATGCG
GAAACGGAAAAATTTGAAATGAAAATAGCAATAATTATTAAGCTTATATCAAATTTATTA
ATGATGAAAGCTCAATATAATGATTTAATAAGTATATTCATGATTTCCAATAATAGTATATTAT
TTTGAGTTTGTAGATACTTGAAATGCAA
```

So, I was able to find a few CRISPR regions. The next step is to find similar matches between the variable spacer sequences (unhighlighted regions) and potential phage genomes. Exact matches may occur, but may also be unlikely due to rapid mutation of phages.

I implemented the following for *Nitrosomonas europaea*, but repeated the process for all of the variable spacer sequences :



I did not take the entire sequence, but rather the middle 20-or-so nucleotides. I did this because nucleotides outside the repeated regions may be associated with the CRISPR structure instead of collected foreign DNA^[6].

However, this did not yield any matches with *Nitrosomonas europaea*.

Next, I did the same for *C. tetani*. A potential match was found.

Finding this sequence in a genome the size of the phage's is very unlikely.

AATGAAGTCAAATAATAACATACCATTTTGTGCTC
AATAATAACATACCATTTT

Chance of occurring randomly in genome the size of
organism's: 9.18×10^{-7}

Finding this sequence randomly in a genome the size of the phage's is very unlikely. It should be noted that other matches were found, but many were based on weaker matches that could have easily happened by chance.

However, according to biobike this is a phage that infects *Prochlorococcus*, which is a cyanobacterium that lives in the ocean. As *C. tetani* is terrestrial, it's unlikely the two came in contact with each other. The specific sequence matches only this phage, and is located towards the end of a Phytanoyl-CoA dioxygenase gene. So, the reason for this match isn't very clear, and could actually be due to coincidence (1 in 90 million can still occur).

Next I looked back at *Mycobacterium-tuberculosis-H37Ra*. I noticed that this bacterium actually had associated phages that had available genomes in Biobike. I decided I should focus on this organism.

Using the method described above, I got the following four matches (Green are matching nucleotides)

Phage - ATCGACGCGAACCTGTCTGACGCG

Bacterium-TTCGAGCCGAACCCGGTGACGCG

Chance of occurring randomly: 4×10^{-6}

Phage: *Mycobacterium*-phage-KBG

This is promising, as the phage is known to specifically infect mycobacteria.

Phage- GGAAGAGATCACGAATCCGGC

Phage- GGAACAGATCAGCCAACGGAC

Bacterium- GGACGTGATCAACGATCCGGC

Chance of occurring randomly: 2.5×10^{-4}

These are Mycobacterium-phage-Kamiyu and Mycobacterium-phage-Tweety. This is good, as these both supposedly infect mycobacteria as well. Additionally, the differences seen are whole-codon changes, which is also promising (as frame-shifts are damaging but codon switches can have much smaller effects).

I also found;

Bacterium- ACGAGGGCGCGGGGCACCCG

Phage- AAAAGGGCGCGGGGCAACCG

(Mycobacterium-phage-kikipoo)

Probability of occurring randomly in phage genome: 1×10^{-6}

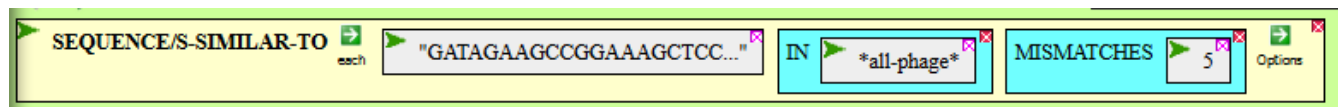
This was a very similar sequence, again occurring in a bacteriophage that attacks the genus of the bacteria whose CRISPR is being analyzed.

I decided to search some more.

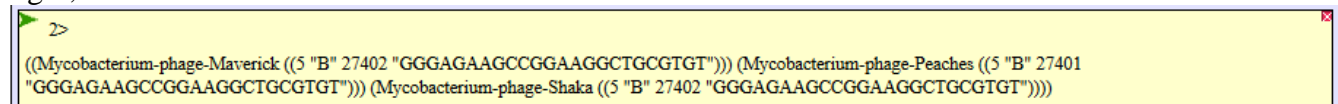
Taking the 24 bp sequence

GATAGAAGCCGGAAAGCTCCGTGC

And searched while allowing for 5 mismatches.



I got;



Mycobacterium-phage-Maverick

Mycobacterium-phage-Peaches

Mycobacterium-phage-Shaka

Which all have the same matching string

Bacterium- GATAGAAGCCGGAAAGCTCCGTGC

Phage -GGGAGAAGCCGGAAAGCTCCGTGT

```

27701 GACGAGTTTCGGGCTACCGGATACAACGGAGCCCCTGCTGGCCGCTCCTGGTTGCGGGACGT
27641 GTCCACGTCGAACGAGTAACGCCAGACCTGGCGTCGATAGTTACAGCTCGGGAGGGACAC
27581 GTTGTGTGCGAGTACATGCAGAGGCTAACGCGCTTCTCTACTGCGATAGAGAGGACCTTC
27521 GCGGAGCTGCTCTTACATCACCAGAGCCCCTGCGGAGACTGCTCGAAGCTGATCGACG
27461 CCGCCGGCATCGAGCGAGTGGTGTACCCGTTTCGAGTGGGAGAAGCCGGAGGCTGCGTGT conserved hypothetical protein
27401 GTGCTGGGCCACCGACCCGTAACACGGGTGGTGCATCCGTGCAGCAAGCTCCCGAAGA Maverick.Mave-0036 (27376 -> 27900)
27341 TCGTGTGCAAGAGCAGGACCCGCTGGGAGCGGAACGAGTACCCGTCGGATGGAACCCG Maverick.Mave-0035 (27211 <- 27405)

```

These are all located inside a hypothetical protein.

I did this again with:

GTTACCCACGCCGATTTACTGGCC

```

s>
((Mycobacterium-phage-Vix ((5 "F" 26138 "GTTGCCGACGACGATCTTCTGGCC")))(Mycobacterium-phage-JHC117 ((5 "F" 26054
"GTTGCCGACGACGATCTTCTGGCC")))(Mycobacterium-phage-Pukovnik ((5 "F" 45895 "GCTACCCACGCGCATATACTCGCC")))(Mycobacterium-phage-Rockstar ((5
"F" 25324 "GTTGCCGACGACGATCTTCTGGCC")))(Mycobacterium-phage-Microwolf ((5 "F" 26057 "GTTGCCGACGACGATCTTCTGGCC"))))

```

Again, all mycobacteriophages.

Mycobacterium-phage-Vix

Mycobacterium-phage-JHC117

Mycobacterium-phage-Pukovnik (with a slightly different matching sequence)

Mycobacterium-phage-Rockstar

Mycobacterium-phage-Microwolf

Phages- **GTTGCCGACGACGATCTTCTGGCC**

Phage Pukovnik- **GCTACCCACGCGCATATACTCGCC**

Bacterium- **GTTACCCACGCCGATTTACTGGCC**

For the phage vix, the match is seen here-

[Select org/contig](#) [Start](#) [Prev](#) [Next](#) [End](#)

```

26078 AACCATCGCCGCGACGTTGTGGCGGCACGGTCACCGGACGCTTCGAGCGGACGGTCTTGGT Integrase
26138 GTTGCCGACGACGATCTTCTGGCCGACGCGGGCCGCGCCCCGGCGCACCCGGAACCTTCAT
26198 CGTCACGCCGTCGTCATGATGTCCTTCCGGCGAATCTCGATCAGCTCCCCGAACCCGAG
26258 GCTCGTCCACGCGAGGAGGTAGACGGCCACCCGGTAGTGCTCATGGACCTCGGCTGCGAC
26318 CGTCTCCAATTCCTGGGGCGTGAGAGCCGCCACGTCGCGCTCGTGGGGGCCCTTCTGCTC
26378 GATCCGGCACGGGTCTCCGAGAGCAGCTTGTCTCCGACGGCGGTGTTTCATGACGGCCCG
26438 GAGGATGTTGTAGGCGTGTCGACGCGCTGTCGGGTGCTGATGCCCCATGCCGGCCACCA
26498 CGCCCGGACGAGGGCTGGCGTCATCTCGGCGACCGCGTGTGCCCCAGCACCGGGTAGAT
26558 CCGCTTCTGGCGTGGGTCTTGTACAGCTCCCGCGTCCCCTCCGCGAGGTGCGGCTCAGT

```

Located at an integrase.

For JHC117,

```

25994 AACCATCGCCGCGACGTTGTGGCGGCACGGTCACCGGACGCTTCGAGCGGACGGTCTTGGT Integrase
26054 GTTGCCGACGACGATCTTCTGGCCGACGCGGGCCGCGCCCCGGCGCACCCGGAACCTTCAT
26114 CGTCACGCCGTCGTCATGATGTCCTTCCGGCGAATCTCGATCAGCTCCCCGAACCCGAG
26174 GCTCGTCCACGCGAGGAGGTAGACGGCCACCCGGTAGTGCTCATGGACCTCGGCTGCGAC
26234 CGTCTCCAATTCCTGGGGCGTGAGAGCCGCCACGTCGCGCTCGTGGGGGCCCTTCTGCTC
26294 GATCCGGCACGGGTCTCCGAGAGCAGCTTGTCTCCGACGGCGGTGTTTCATGACGGCCCG
26354 GAGGATGTTGTAGGCGTGTCGACGCGCTGTCGGGTGCTGATGCCCCATGCCGGCCACCA
26414 CGCCCGGACGAGGGCTGGCGTCATCTCGGCGACCGCGTGTGCCCCAGCACCGGGTAGAT
26474 CCGCTTCTGGCGTGGGTCTTGTACAGCTCCCGCGTCCCCTCCGCGAGGTGCGGCTCAGT

```

Also at the integrase protein.

Rockstar:

```
25264 GACCAITCTGIGCGACGTGTGGCGGCACAGTCACCGGCCGCTTCGACCGGACGGTCTTGGT Integrase
25324 GTTGGCCGACGACGATCTTCTGGCCACGCGGGCCGCGCCCGGGCGCACCCGGAACAGCAT
25384 CGTCTCCCGCTCGTCTGTATGTCCTTGCGGCGAAGCTCGATCAGCTCCCCAAACCGCAG
25444 GCTCGTCCAGGCGAGAATGTAGACCGCGACCCGGTAGTGCTCATGCACCTCGGCCGCGAC
25504 GATCTCCAGTCTCTCCGGAGTCAGGGCTCCACGTCGCGCTCAGCCTCTGCCTTGATCTC
25564 GATCCGGCACGGGTTCTCCGACAGCAGCTTGTGTCACGGCGGCTTGGCAGACGGCGTG
25624 GAACATCCGGTACGTCTGTCTCCTGGCGGCATGTGCTTGTGTCACCTCCGGCGAACCA
25684 CGCACGGATGAGAGCCGGTGTAGCTCGGCCACCGGGTCTCGCCAGGACCGGGTTGAT
25744 GCGCTCCGCGCATGAATCTTGTACAGCTCTCGGGTGCCTCAGCGAGCGGCCGCTCCTC
25804 GATCCACTTCCCGGTACTCCTCGACGGTGATGGACGAGGCTTGGGCTTCTTCGCCCG
25864 TTCCTCGGGCGGCTCCAGGCTCCATCTCGATGAGCCGCTTCTCCTGGCCAGCCAGGC
25924 TTCGGGTCATCCGGTGTGTCGATAGGTGTGACGGCGTAGTACCGCACCCCGTCCAGCGG
25984 GTGGACATACGAGGCTTGGATACGCCCACTCCGCATCGTCTTCAGTGATCCCCATGACCG
26044 TCGTGAGGCTGCCACCGAGGCTCCTTTCTCCCGTCAGAAAGGGTACCGAATTTGCAACTC
26104 TCATGCAACTCCCGAGGCTCATCCGTTTTACGACCTGCAATTTCTTTCACTTTGAGAG
26164 TTGGATCTGGCGAGGTTAAAACCTGCTCTGACCTGCACATACAGTCTGATACGGGCTCT
```

Also at itegrase.

Microwolf:

```
25997 GACCAITCGCCGCGACGTGTGGCGGCACGGTACCGGACGCTTCGAGCGGACGGTCTTGGT Integrase
26057 GTTGGCCGACGACGATCTTCTGGCCGACGCGGGCCGCGCCCGGGCGCACCCGGAACACTCAT
26117 CGTCACGCGCTCGTCCATGATGTCCTTCCGGCGAATCTCGATCAGCTCCCCGAACCGCAG
26177 GCTCGTCCACGCGAGGAGGTAGACGCCACCCGGTAGTGCTCATGGACCTCGGCTGCGAC
26237 CGTCTCCAATTCCTGGGCGGTGAGAGCCTCCACGTCGCGCTCGTTGGGGGCTTCTGCTC
26297 GATCCGGCACGGGTTCTCCGAGAGCAGCTTGTCTCGACGGCGGTGTTTCATGACGGCCCG
26357 GAGGATGTTGTAGGCGTGTGACGCGCTGTGCGGGTGTGATGCCCATGCCGGCCACCA
26417 CGCCCGGACGAGGGTGGCGTCATCTCGGCGACCGCGTGTGCGCCAGCACCGGGTAGAT
26477 CCGCTTCTGGCGTGGGCTTGTACAGCTCCCGCGTCCCTCCCGGAGGTGCGGCTCAGT
26537 GAGCCACTTCTTGGTGTACTCCTCGACCGTGATGGAGGATGCGGCCTTCTTCTGGCCCG
26597 CTCGGCGGCGGCGTCCAGGCTCCATCTCGATGAGACGGCGCTCGGAATTTAGCCACGC
26657 TTCGGCTCCATCCGGTGTGTCGATAGGTCTGCACTGCGTAGTACCGCACGGCCGTCAGTGG
26717 GTGGACGTACGACGCTTGGATACGCCCGCTCCGCATCGTCTTCAGCGCTCCCCAGGAGCG
```

Also at integrase

For Pukovnik, it was located at a completely different egion in a different protein.

```
45835 GCGGAGTCTCTCCACCAGCAGCAGCGGCTTCTTCTGTCCTCGGACATCGTCTTGAA Phage repressor # Pham54
45895 GCTACCCACGCGCATATACTCGCCGTGGTCGGAAGCCTCTGGTAAGCCTTCGACTTCCC
45955 GTGCAGCTTAGTGGTTTTCCACGGCCACGCTCTTGGACGATCGCTAGTGGTCAATCG
46015 TCCCCCGTAGTCTTCTTCTGCCACGAAACAGCCTGGCGGGTGACGCCATGCATGTCGCG
46075 GATCTCGCTCTGATGAACCCCTTCTGCGAAGATCCTCAATCGTCTGAGGGTCAAAGG
46135 CTGTCGCGACGGGGCCGATCGCTCGCCACGTGTGATTTGCGGCTCATGTTCCCTC
46195 CATGAGAAAGGTTCAAGTTGATTTCTCCTGTCAAGGAGAATGTAGGTGACTGTCAAGTCA
46255 ATCTCTTCCCATAACTCGTGCCTTCGACCGGCTCTCGATCTCAGACCTTCGGCTCT
46315 TACCAGGTGCTGCTAGTAGCTAGCTGAACAAGGCTTACGCACTCGTAAACCTAGCTGGT
```

A phage-repressor gene.

It is interesting that both viruses contain the matching sequence yet do not appear to be related. Still, it is reasonable to assume that

Vix, JHC117, Rockstar and Microwolf are all related to each other closely, as they show highly homologous DNA sequences.

I again searched with:

TGCGCGCACAAACGCATCCGCCATCCA

One match:

```
17> ((Mycobacterium-phage-Wildcat ((5 "B" 36894 "TCCGCGCGCAACACATCAGCCAGCCA"))))
```

Again, a mycobacterium phage.

Bacterium- TCGCGCACAAACGCATCCGCCATCCA

Phage- TCCGCGCGCAACACATCAGCCAGCCA

```
37373 AATAGCGCCAGGGTCCCCTTGGCCCTGCGCGCGCTCGCCCCACTCACGGTGGGAGATA
37313 CCTATCGGCACGCAACCCAGCTTGGGAGAATAGCCGCATGCACCTTCACAGTTGCGTC
37253 GTACTGCACATCTGGCCAGCCCTCACGGTGGGAGCATCCTTGGGAGGCAGGATCGCCAC
37193 CTCGACGCCGATCATCACCGGGTTCGCGTTGTTGTAGGCAGGCCAGGATAATCGCCACG
37133 ACCAGCGTGGTTCGCTTACCGATAACCACACCCACACGCCTCGCGTTGGCTTGATCAG
37073 AATGTGGGCAGCCAAACCCAAAGTCGGGTGGAAAGCAATGCCCTCCGGAGTCTCGTTCCG
37013 AGAACAGTGTGGTGAACATGACCCCCAGAAAGTTCCTGGTTCGCGCTGGCCAGGGT
36953 CTGCCAACCATCCACCTCGAACACGCGAAGACCCCGCGCGCAACACATCAGCCAGCCA
36893 CCACGGTTCACAGTGAACCCAGGCTCAGGAGTAGTGGCAGGAGTGTAGGACCCATCAA Phage endolysin
36833 CGGGATGGCCAGCACGCGATTCCATCGAGCCTGACGATCCGTCACGCCATTCCGAAGCTG Wildcat.Wildcatp49 (36288 -> 37796)
```

This is a page endolyain gene.

More searching with:

GGGGACTGTGGACGAGTTCGCG

```
19>
((Mycobacterium-phage-DS6A ((5 "F" 35433 "GGGTGACGGTGGAGGAGTTCGCG")) (Mycobacterium-phage-Pari ((5 "B" 37214
"GGGGACCAAGGGCGAGTTCGAG")) (Mycobacterium-phage-Backyardigan ((5 "F" 31589 "GGGAGACGGTGGATGGGTTACG"))))
```

Three matches. All mycobacteria. DS6A, Pari, Backyardigan. All slightly different.

Bacterium: GGGGACTGTGGACGAGTTCGCG

DS6A: GGGTACGGTGGAGGAGTTCGCG

Pari: GGGGACCAAGGGCGAGTTCGAG

Backyardigan: GGGAGACGGTGGATGGGTTACG

```
35373 TGCACGTGAAGCTGACGGCGGCACGCTGACCGAGTACCACGGGTGCTGCGCGGCGG hypothetical protein
35433 GGGTACGGTGGAGGAGTTCGCGGGGAGGCGATTGCGGAGCGGTTCCGGTGTGGTGA
35493 AGGCGCAGGCGGGCGGGCGTCCGGGCGGTAAGCGGTGGCATGATGGCCGCCACATT DS6A.DS6A-0051 (35536 -> 36210)
35553 GGACCGCCGGCAGTGCCTGGGGTGTGGCGAGGGCTGCCGGCCGACGCCACCCGCTGCG hypothetical protein
35613 GAAGTGGTGCAGTGCAGCGGTGCCATTCTCGGGCACGAGACGCCCGGCTGATCGGGAGC
35673 GAAACGTAAGGCTCGGAAGCGGGAACGGCACCGGGAGCGCATGGCGACGGATCAGGCATA
35733 CAAGGATCGACCCGGGCAAGAGCGGGCGCAGAAGCAGCGGCGAGCGGGCGGCGAAAGT
35793 GGCCGCCATGCCGCTGCGGCAGTGCGCATGGTGGCGGAGTTCGTTGGGGCCGGCGCCCC
35853 CTGGTCTATCGGATATGCTGCGGGCTGCGTATCGGACTACCGGAGCGCAATCA
```

In DS6A this occurs in a hypothetical protein

pari, this also occurs in a hypothetical protein.

```

37693 ACGACATCCGGCACGTGCCGGATGGTGTGACGTAATCGACAGGTTCCGGGAACCGGCAAT
37633 GCTCGCGGCGGTGATAAGCCGGTTTAGAGCAGAGTTCGTGAACGAATGGGCTCCGTTCA
37573 CCGAGGTCGTGCAATGATCGTGGTAAGAGCCAGACCCTCCCGAGGGGATTCCGCGTCT
37513 CCGAGTCGAGCTCGGCGGTAAACCGTACCTACACAACAGTCCGCAACCGGATCAACCGCT
37453 GGCAGCGCTACCGATCAGTACGCGCGGAGCTATCCCTCCTCCGAGAGAAGGGAACCAA
37393 ATGACTAAACGAATCGTCTTTCTACCCGACACTCAGTTGCCCTTTCGAGGCGCGCAAGGAG
37333 ATGCAGGCGGTCAATCCGCTTCAATCGGGGATGTCCAGCGTACGGCGTGGTACATATCGGT
37273 GACGTCCTAGACCTGCCCGACGCCCTCGCGCTGGAATCGGGGACCAAGGGCGAGTTCGAC
37213 GGTTCGGTGTACCGCGACCGGACTACGCCAAGAAGAACCTGATGGAGCCACTGCGCAAG
37153 GTCTACGACGGCTGGATCGGGATGCACGAGGGCAACCAGATCTGCGAGCCCGGAGTAC

```

Protein of Unknown Function
Pari.Pari-0052 (37560 <- 38216)
Conserved hypothetical Mycobacteriophage
Pari.Pari-0051 (37393 <- 37560)
Conserved hypothetical protein
Pari.Pari-0050 (36629 <- 37393)

in backyardigan,

```

31529 GACAACATGCCGGTGTCTCGATCATGGTGTCTCTACTTGTGGTACTGCCCCCGGACGATCC
31589 GGGAGACGGTGGATGGGTTCACTCGAAGGACGCGGCGACATCGCGCCGGGTGACGCCCG
31649 CTCGGACGAGGTCCTTGATGAACGCGACCTCGGACTTGTGAGCTTCCGGTCCGGTGGGTC
31709 GGTGGGACCCTTGGTCTCCAACCTTGGCTCGCAGCTCGCGGTTCTCCTCCACGAGCCGCT
31769 CGTTCACCGACGCCAAGTGTCTCCGCTGTTGGTACAGCGTGGTGTTCGAGCTTCCGAGAG
31829 CGTCGATGGACGCAATCGCCTCTCGCAGAGCGACTTTCAGTTGCTTCTTTCGATCAGTT
31889 GTCTTCCAGCGGTGACACCCGCTAGTACATGAGGTTTCGGGCGGTAGAAGCTGAGATACGC
31949 TCCGCTGTCCGCCGAGATGACCAACGTTCTCTGATGGGGTTCGACGTTGACCTCACCAC
32009 CGTGGGATGATGGTCCCGTTCGATGAGCAGGACCGTACTTCTTTCATTCGCCCTCT
32069 CAGTAGCTGTAGGGCTCGTTGGGGATGTCTGGTAGGTTGGGAGCGATCCTCCGGAGC
32129 TGCGCGACGAGTTCCTTCCAGTTTACCGGATTTCCGGCTCCGCTGCCCTCGTCCAGCGG
32189 GCCTTGTGACGTAGCGCCAGGCGCGGTGGTTGCCGGTACGACCATCGGTGAGTTGGTC
32249 ATGTTCCGCGAGGACCGCTCGTGTCTCGCTCGCGAGCCTTCTTCCGGGGCAACCAGCCGTC
32309 TCCAGCGACGACGAGCGGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG

```

Backyardigan.Backyardigan-0044 (31560 <- 318
hypothetical protein
Backyardigan.Backyardigan-0045 (31883 <- 320
hypothetical protein
Backyardigan.Backyardigan-0046 (32068 <- 327
Thymidylate synthase thyX (EC 2.1.1.-)

Another hypothetical protein.

Another search:

CGCGGGAAGAGATCACGAATCCGG

```

31>
((Mycobacterium-phage-Gadget ((5 "B" 30187 "CGCGGACGTGATCAGGGATCCGG")))) (Mycobacterium-phage-Akoma ((5 "B" 30181 "CGCGGACGTGATCAGGGATCCGG"))))

```

Two more mycobacteriophages.

Bacterium: **CGCGGGAAGAGATCACGAATCCGG**

Gadget and Akoma: **CGCGGACGTGATCAGGGATCCGG**

```

30846 ACTCCTTGACGAAGTAGATCGTGTTCGGTCTTCGACGTGATCGCCGCTACTCCGCG
30786 CCGTGCCCTTCCAGATGTTAGGTTGGTCAGGCCGACCGGTGTAACCGGCCACGACGC
30726 CGGACCCTGCGAGGATCCGCCCGGTACATCGCCACGTTGACCGTGCCTCGAGT
30666 AGTGTTCGCCCTTCTCGACGCCGACACCCGAGTACCGGTAGTCCGATCCGTAGAGCGACA
30606 CCGGGGTCGAGTCGACGCGTGCAGTGGACCTTCGTGCCGTTGACGCGGGCCTCGAAGTACC
30546 TGGTCTCCGCCAGGTGCCGCCCTTGAACGAGATCGAGCATCCGCCGTCAGCACGTCGG
30486 CGCCGAGAGAGTCCGTCCTACTGCGTAAACGTTGCCGCTGCGTAGCTGAAGAACCGGA
30426 TCTCGTCCCACGCGTATCGGGCCGCGACGAAGGCGGTGCCGTTGTTGACCCGCGCA
30366 TGAGCCACAGTAGTTGGAACCTGCTGCGCCGAACCAACCGTGCAGGGGACGAGCGAA
30306 GCACCGCGCTCACCTCGAACAGGTCGGTGGCAGGGGACCGCGTTGAACAGGTAGAACT
30246 CGCGTCCGGCCGACGAGCCGACAGGCGAGCTTTCGCGGACGTGATCAGGGATCCGG
30186 ATCCGGTGTGATGACCTTGGTGAACATGGACGCGGAGTGGACCGTCCACGAACCTCAC
30126 CGATGTTGACGACCTCGGAGGACGACGCGGGTGCAGGCCGCCACCTGCGCTGGAGGT

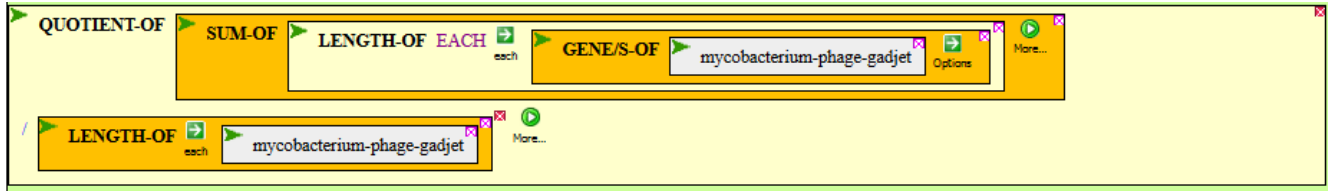
```

Phage protein
Gadget.Gadget-0031 (28620 -> 30848)

Located on a phage protein.

Out of the 8 phages looked at so far, all matching items have been located on a gene.

To see if this was odd or not, I decided to look at what roportion of the paghes' genomes were coding regions.

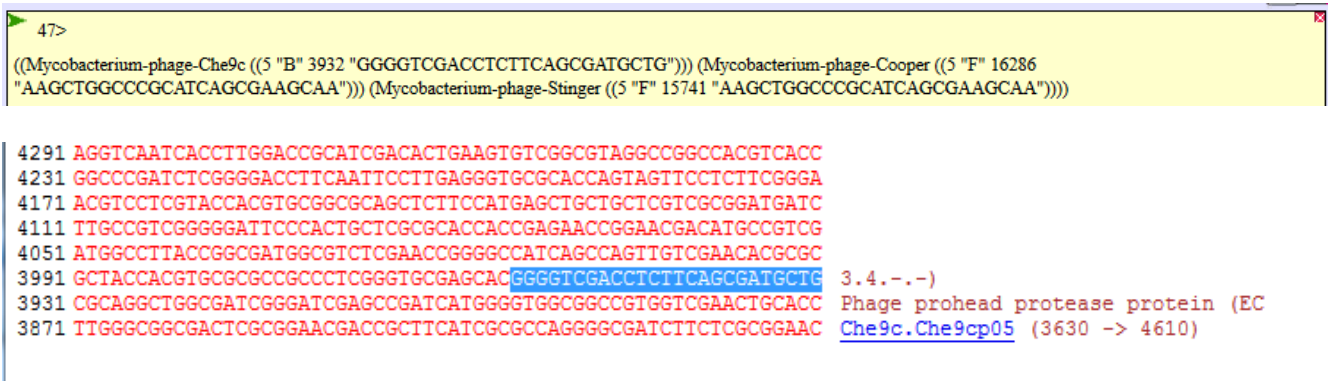


- Gadget – 95%
- DS6A - 94%
- Pari- 90%
- Backyardigan – 91.8%
- Wildcat- 92.2%
- Vix – 89.9%
- Pukovnik – 93.1%
- Maverick – 91.9%

Average: 92.2% coverage.

So, there's about a fifty percent chance (92.2^8) that I would find the sequences exclusively on genes.
More searching.

GAGCTGGACCGCATCAGCGATGCTG



Prohead protein. Coding region: 94.7%

```

15681 ACGGTGGCCCGCCAGCCGAGCCCGCCGCCACCACCGCACCAGCCCGCGATCCCGT
15741 AAGCTGGCCCGCATCAGCGAAGCAA CAAACCGCCCCGACCTTGTGGCCGGGGCGGTTTGT
15801 TGICTGGGGCTTAGGGCGGCTGCGGCCACCGCCTCGTTCACCAGGTGCTAGGTGAGCTGGC
15861 CGITGCCGTAGAGCGGGCCACCTGGCTGGGCTGGAGGGTGAAGCCGGGCAGCGGCACCGA
15921 GGGCGACGGTGACGGCGGGTGTATGTCGCGGCAGTAGGTGCTGGTGACCTCCACGGCGT
15981 GGACATCAAAAGGGCGCGGGCCAGGCGAACCTCGTCAGCGGTGGTGGCACCGTGGGGCGA
16041 TGGCCAAGCGGAACTCGCGGTGATGGCCTGGGCGGTGAACGGGCTGATGGTGGACATTT
16101 TGAATCCCTTACCTGCGGGCGGGGGCGGTCCCGCCTTGTGTGATGTGAACACACTAACCC
16161 GCCTTAGTCGGGTAAAGTCAAGATGACCGATAACCGCACCGCCACCGGATACCCTGGCGTGG

```

For the mycobacteria phage stringer, the match actually fell (mostly) between two declared genes. Coding region is 94.9%.

More searching...

Another match:

```

152857 TGCTGCCCGCAACGAAAGGGGATGAAGTCATCTTCGAGAGATTCACCGACCGCGCCCGTC
152917 GCGTCTGCGTCTCGCCCAAGAGGAAGCGCGCATGCTCAACCATGCGTACATCGGCACCG
152977 AGCATCTGCTGCTCGGGCTGGTCCACGAGGGCGAGGGCGTAGCCGCCAAGGCGCTGCAGG
153037 CGCTCGACATCAACCTGGAGGAGGTGCGTGCCGAGGTGAGGAGATCATCGGCCACGGCC
153097 AGCAGGCACCCACCGGCCACATCCCGTTCACCGACCGCTCCAGCGCGTCTGGAGCTGT
153157 CGCTGCGCGAGGGCGCTGCAGCTCGGCCACAACACTACATCGGCACCGAGCATCTCCTGCTGG
153217 CCCTGATCCGCGAGGGTGGGGCTGGGCTGCCAGGTGCTGGTGAAGCGCGCGCGGAGC
153277 TGACCAAGTCCGCCAGGTGGTTCATCCAGTGTCTGTCGGGTATGACCCCGAGCTGGCGA
153337 AGAAGCGCAACCGGGAGGAGGAGCAGACCACCATCCAGGGTTCAGCTCAACGACACCGGAGC
153397 GTGTGCTCGAGAAGGTCAACGGCAAGATCGAGAAGCTCTCGACCACCAACCGGAGGCGA
153457 TCATCGGCAACGTGTACGTGGTGACCAAGATCCTCAACGACCGGATCGAGCCGCGCG
153517 TCGACCTCGTAGCAGCACTCATCGAATGGAAGGGAAATCGCAGTATGACGGCCAAGAGCGC
153577 GAGTCAGATCATCCCGGCAAGCTGGTGGACTTCATCCAGCGCGAGGGAGGTGCCCTGGC

```

94.9%

So, there was one instance of a non-gene match and 10 instances of gene matches. This is 91% actual instances of the match landing on a gene. The average gene coverage for the viruses so far is 92.9%. This is reasonably close to what would be expected.

Clearly more data could be obtained and would likely be beneficial in elucidating a further pattern. However, based on the obtained results, it seems reasonable to conclude that there is a fair chance there is no correlation between coding/noncoding viral DNA and the variable spacer regions in CRISPRs. Still, it was interesting to find a number of matches to known mycobacterial phages, as this likely indicates previous contact in the organisms evolutionary history.

I think I was able to answer my question, at least in a preliminary sense in that a lot more research would be required to really answer the question but this was a decent start. I found very little relationship between coding vs non coding regions in phages and CRISPR regions in bacteria. I was also able to identify quite a few phages that likely infect one particular mycobacteria. Some of the phages seemed related, others didn't, and some even contained highly similar sequences even though they otherwise seemed very much unrelated (what was very surprising – perhaps they picked up similar DNA during a lysogenic phase, or some other mechanism for horizontal gene transfer).

Future research into this would benefit from further automation of the process, although quite a few of the key programs I used took between 10-50 seconds for biobike to execute. Further automation may take long periods of time for the computer to run. Still, more data could construct a better picture and determine if there is in fact a relationship between selection of variable sequences and the region of the phage genome it comes from. This could actually be useful, as CRISPRs are rapidly being integrated

into biotechnology purposes, and further understanding could benefit the field.

References

- [1] Hans-W Ackermann. 2011. Bacteriophage taxonomy. Under the microscope. 90-94
<http://journals.cambridge.com.au/UserDir/CambridgeJournal/Articles/11%20ackermann244.pdf>
- [2] Wommack, K. E.; Colwell, R. R. 2000. Virioplankton: Viruses in Aquatic Ecosystems. *Microbiology and Molecular Biology Reviews* 64(1): 69–114
- [3] Horvath, P., Barrangou, R. January 2010. CRISPR/Cas, the Immune System of Bacteria and Archaea. *Science*. 327(5962): 167-170
- [4] Haft, D. H.; Selengut, J.; Mongodin, E. F.; Nelson, K. E. 2005. A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. *PLoS Computational Biology* 1(6): e60
- [5] Barrangou, R.; Fremaux, C.; Deveau, H.; Richards, M.; Boyaval, P.; Moineau, S.; Romero, D. A.; Horvath, P. (2007). CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* 315(5819): 1709–1712.
- [6] Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, et al. 2008. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *Journal of Bacteriol* 190: 1401–1412