Characterization of palindrome usage in mycobacteriophage genomes

Julia Gross, Kenyon Alexander, Brandon Le, Zachary Ricca, Hwai-Ray Tung, Benjamin Siranosian, Yang Zhou, Sarah Taylor, Christopher de Graffenried
Brown University, Providence RI

**Background**

Palindromes: DNA sequences with identical reverse complements.

The reverse complement of GATC is GATC!.

We characterized the presence of palindromic sequences in length 4 and 6 in mycobacteriophage genomes to look for global trends and specific interesting examples.

**Methods**

Palindrome counting: all length-4 and 6 palindromes

Example using a sliding window:

GATGATGATCATG

<table>
<thead>
<tr>
<th>Palindrome usage in cluster B phage</th>
<th>TUD value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATCG GACG</td>
<td>1</td>
</tr>
<tr>
<td>GATT GATC</td>
<td>1</td>
</tr>
<tr>
<td>ATCA TCA</td>
<td>1</td>
</tr>
<tr>
<td>GATC TGC</td>
<td>1</td>
</tr>
</tbody>
</table>

Normalization of counts

Palindrome occurrence depends on the underlying nucleotide counts in the genome. To normalize, we computed the expected number of a palindrome in a genome with identical nucleotide composition but randomized sequence order:

\[ \text{Expected} = \left( \frac{4}{A^4} + \frac{4}{C^4} + \frac{4}{G^4} + \frac{4}{T^4} \right) \times N \]

where:

- \( A, C, G, T \) = genomic frequency of respective nucleotides
- \( A_{pc}, C_{pc}, G_{pc}, T_{pc} \) = tetranucleotide frequency of nucleotides
- \( N \) = length of genome

The normalized count is the observed count divided by the expected count:

\[ \text{TUD} = \frac{\text{Observed}}{\text{Expected}} \]

This removes bias from GC count and highlights interesting patterns of palindrome occurrences. One TUD value for every possible palindrome. We calculated this in all mycobacteriophage genomes.

**Palindrome Q&A**

Q1: Are palindromes globally avoided in mycobacteriophage genomes?

Restriction enzymes can cut palindromes in double-stranded DNA. Are there systems that recognize palindromes in Mycobacteria?

Q2: What influences palindrome occurrence in phages?

We searched the restriction enzyme database REBASE for information on restriction/modification and methylation systems in Mycobacteria.

Q3: What palindromes are underrepresented?

Some palindromes are used infrequently or not at all in some genomes.

R3: Certain adenine-thymine palindromes in A-cluster phages

- Not in any A4 genome
- Not in half of A2 genomes

Conclusion:

Something is suppressing ATAT and AATT usage in cluster A phages. Maybe there are targets for currently unknown restriction enzymes in the native host bacteria of cluster A phages?

Q4: How did GATC become so frequent in B3 genomes?

GATC is used 4 times more frequently than expected in B3 genomes. We investigated if consistent single nucleotide polymorphisms (SNPs) lead to an abundance of this palindrome.

R4: Interesting mutation patterns

- B2 phages maintain palindromic GATC sites
- Related phages commonly have NATC

**References**

2. (2007).

**Acknowledgments**

We would like to thank:

- The other instructor of the Brown University Phage Hunters, Peter Shank.
- The MDI Lab staff at Brown, and HHMI SEA-PHAGES for providing us with the opportunity to conduct this research.
- Chen Ye for designing some of the graphics.