Abstract

With recent advancements in DNA sequencing technology, there has been a rapid growth in the number of available sequenced bacterial genomes. Embedded in these genomes is a multitude of foreign DNA, including transposons, insertion elements, as well as bacteriophage DNA known as prophage. While there are fewer fully sequenced phage genomes, these newly available bacterial genomes highlight phage genomics in the context of temperate phages. We present a comprehensive survey of 11,941 bacterial genomes, which were scanned for prophage regions usingPhiSpy, a weighted algorithm used to identify prophage regions using a variety of characteristics of phage. A total of 67,022 prophages were identified, with a mean length of 23,351 bp. The data is being used to develop novel methods of identifying prophage insertions into tRNA and bacterial genes to reveal that phages do not preferentially insert into tRNA sites as previously believed. A phage gene heat-map demonstrates the favored position of hallmark phage genes in relation to the position of the integrase gene. Altogether these findings provide a new perspective on the mysterious nature of phage behavior - which has significant implications in understanding trends in microbial ecology as well as bacterial gains of function such as antimicrobial resistance.

Approach

<table>
<thead>
<tr>
<th>Bacterial SEED genomes</th>
<th>Integrate alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prophages</td>
<td>Condensed Dataset</td>
</tr>
<tr>
<td>Prophage Statistics</td>
<td>Gene heatmap</td>
</tr>
<tr>
<td>• Prophage length</td>
<td></td>
</tr>
<tr>
<td>• Species distribution</td>
<td></td>
</tr>
<tr>
<td>• Codon usage bias</td>
<td></td>
</tr>
<tr>
<td>• Hypothetical regions</td>
<td></td>
</tr>
<tr>
<td>Prophage Insertions</td>
<td></td>
</tr>
<tr>
<td>tRNA</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td></td>
</tr>
</tbody>
</table>

Future Directions

1. Release Insertion Site Analysis Package
2. Improve condensed dataset for heat-map
3. Make correlation of insertion sites with Integrase proteins
4. Integrase domain consistency

Acknowledgements

I thank Dr. Anca Segall and Dr. Barbara Bailey for their constant insight. Daniel Cuevas and Genivaldo Silva for their continued support. NSF DEB 1046413 for funding this project.

For additional information and inquiries, please contact me at: hkang408@gmail.com