|  |  |
| --- | --- |
|  | **Power**  Florida Gulf Coast University  May 13, 2014 |

We are confident with Power’s annotation and our start site choices yielded many query 1 to subject 1 matches with A2 phage in NCBI and PhagesDB. In summary: (1) synteny was followed with the structural genes, (2) there appear to be no new genes, (3) no t-RNAs were discovered, (4) there is a frameshift in the tail assembly chaperone, (5) there were three questionable orfs, and (6) several large gaps were identified.

**Frameshift**: 15725-16132 and 16198-16557 for tail assembly chaperone. The slippery sequence involved was GGGGGAA. This orf bore a high similarity in protein sequence with Turbido and a Phamerator map confirmed this similarity. The frameshift was annotated by reading the first G in the slippery sequence as two, which allowed the gene to shift up one reading frame at Glycine at 16105 bp.

**ORFs:**

(1) We **removed** 26415-26537 orf, located after the integrase gene. It was only called by Glimmer with a strength of 1.26. There were no matches in PhagesDB BLASTp. In NCBI BLASTp, there was only one match to a hypothetical protein *Bacillus sp*. 37MA, 50% identity, 53% coverage, E-value=5.4, query 8 to subject 34. There was no coding potential in the GeneMark graphical output with *M. smegmatis* preferences. In HHpred, the best hit was to conotoxin Y-PIIIE; neurotoxin, acetylcholine from *Conus purpurascens*, probab=79.47, E-value=0.67. And the second hit was to 82 prophage-derived uncharacterized protein YBCO from *E. coli*, probab=73.39, E-value=1.3.

(2) We **didn’t add** the following feature, but debated whether it should be added, 25148-25312. It was not called by DNAMaster, but the GeneMark graphical output using *M. smegmatis* preferences showed strong coding potential in reading frame +2. It had an SD score of 567 and a length of 165 bp. The problem is that if added, this orf would create a 85 bp overlap with the integrase gene. NCBI BLASTp had only 2 matches, both with poor E-values (>5.0).

(3) We **left in** the following feature 51750-51827 orf. However, the orf is very small, only 78 bp, and has a very tiny potential in the GeneMark graphical output using *M. smegmatis* preferences. We left it in, but we’re not sure if it’s real.

**Gaps**: There is a very large gap from position 52152 bp to 330 bp that encompasses the 3’ end of the genome and the beginning of the 5’ end. Since Power is 53395 bp in length, the gap is 1573 bp. The gap was BLASTed in NCBI BLASTx and no good matches were found. All E-values were > 6.9. Additional large gaps were identified: between gp1 and gp2 (217 bp), gp 2 and gp 3 (142 bp), gp7 and gp8 (111 bp), gp 23 and gp24 (110 bp), gp 32 and gp 33 (393 bp), gp57 and gp58 (209 bp), gp63 and gp64 (131 bp), gp73 and gp74 (241 bp), gp74 and gp75 (133 bp). We did not find coding potential in these gaps.