Case Study: *Cluster A DNA primase*

Phage Genome: *Peaches*

Genome Coordinates: *37497-38351*

Objective: Explore a “gap” in the predicted CDs features in a cluster A genome

determine if a gene is present

determine putative function

1. Collect Evidence: insert screenshots relevant to THIS REGION of the genome as indicated below. of GeneMark, BLAST results, Aragorn, Phamerator, HHPred, choose start window, frames window, etc. Make sure that your evidence is readable and complete.

A: *screenshot of BLASTN results on phagesdb.org*

B: *screenshot of phagesdb.org pham page with functions*

C: *Autoannotation results for this region (what do the DNA Master notes say are the Glimmer and GeneMark results?)*

D: *screenshot of the DNA Master frames window*

E: *screenshot of GeneMarkS output. If no coding potential in region, add GeneMark-host output(s)*

F: *screenshot of regional Phamerator map, with relevant phages*

G: *screenshot of DNA Master “Choose Start” window*

H: *screenshot(s) of DNA Master BLAST results, including alignment tab(s)*

I: *screenshot(s) of Starterator report*

J: *screenshots of HHPred results*

K*: screenshot of Aragorn results*

L*: screenshot of tRNAscanSE results*

M*: screenshot of Phamerator map of entire genome*

N: *other*

2A. Is this a gene? \_\_Y; two genes\_\_

Rationale: Write a one-sentence explanation for each piece evidence above and how it supports your answer to 1, or write NI for not informative or n/a for not applicable.

A: Many A4s show high sequence similarity in this region, so they are likely to have the same gene content.

B: The autoannotated gene is a member of a pham with a lot of other members, and they have been assigned the function DNA primase

C: Only one gene is in the autoannotation, both programs predicted this gene, albeit with different starts.

D: There are no other genes overlapping with the predicted gene, however, there is a gap in this region at the transcriptional start of this gene (towards the 3’ end of the genome)

E: There is coding potential in the predicted gene, but also in the orf upstream that was not predicted. Adding the other ORF would yield a large overlap with the predicted gene.

F: Lots of the final A4 annotations contain both genes, even with the large overlap, suggesting that there is a reason for adding the second gene.

G: n/a for gene content

H: both genes are present in GenBank in multiple other phages

I: n/a for gene content

J: both genes have a good alignment to DNA primase crystal structures over most of their length with a probability in the 90s.

K: n/a

L: n/a

2B. Rank the evidence from most-to-least compelling. include a one -to -two sentence rationale to support your order or write NI for not informative or n/a for not applicable.

J. both genes have high scoring alignments across most of their length to crystal structures in the PDB.

E. There is excellent coding potential in both ORFs, even though adding the upstream one means that there will be a large overlap with the downstream one.

C. Both gene prediction programs predicted the downstream ORF is a gene.

A, B, F and H: the A4s are very similar in nucleotide sequence in this area (A), and lots of other final A4 annotations include both genes with the overlap (B, F, H).

If your answer to 2 is “no”, you are done with this case study.

If not:

3. What is the start coordinate for your gene? \_\_\_\_*38138 and 38367*\_\_\_\_

Refer to evidence in part 1.

Rationale: Write a one-sentence explanation for each piece evidence you’ve included and how it supports your answer to 3 or write NI for not informative or n/a for not applicable.

A: n/a

B: n/a

C: GeneMark selected 38138 as the start

D: 38367 is the longest possible start for the upstream gene. The frames window is not really helpful for the start of the downstream gene.

E: Coding potential really spikes right at 38000, so both the Glimmer and GeneMark starts would include all of it. Only the 38367 start captures all the coding potential for the upstream gene.

F: n/a for start choice

G: Both of the starts listed above have the best RBS scores of the bunch.

H: BLAST alignments are 1:1 for both starts selected above, showing that many of the final annotations use these starts.

I: The start at 38138 is conserved throughout the entire pham, the Glimmer start for this gene is not.

J: Both selected starts yield alignments to crystal structures--- even though there is a large overlap in the CDs in the annotation.

K: n/a

L: n/a

3.B Rank the evidence from most-to-least compelling, include a one-to-two sentence rationale to support your order or write NI for not informative or n/a for not applicable..

J: most compelling—alignments to crystal structures include sequence that can only be obtained by choosing the longest start. The best alignment to the downstream gene indicates that it is missing the first 100aa. The upstream ORF aligns to those missing 100aa.

C: coding potential is only all captured by longest start

I: STarterator for the downstream gene supports the start that results in a big overlap

the rest are less informative, or not informative

4A. What is the function of this gene? (choose from approved list): (ie *terminase, large subunit)*

DNA primase, for both.

4B: Rationale for evidence:

Rationale: Write a one-sentence explanation for each piece evidence you’ve included and how it supports your answer to 4A or write NI for not informative or n/a for not applicable.

A:

B: most members of the pham are labeled DNA primase

C:

D:

E:

F:

G: most close matches are DNA primase

H:

I:

J: Alignment across most of the sequence is to crystal structures of DNA primases

K:

L:

4C: Rank the evidence from most-to-least compelling. Include a one-to-two sentence rationale write NI for not informative or n/a for not applicable.

J: : Alignment across most of the sequence is to crystal structures of DNA primases

B: most members of the pham are labeled “DNA primase”

H: close matches are “DNA primase”

5. How does this function influence your start choice or gene/not a gene decisions?

**a. confirms**

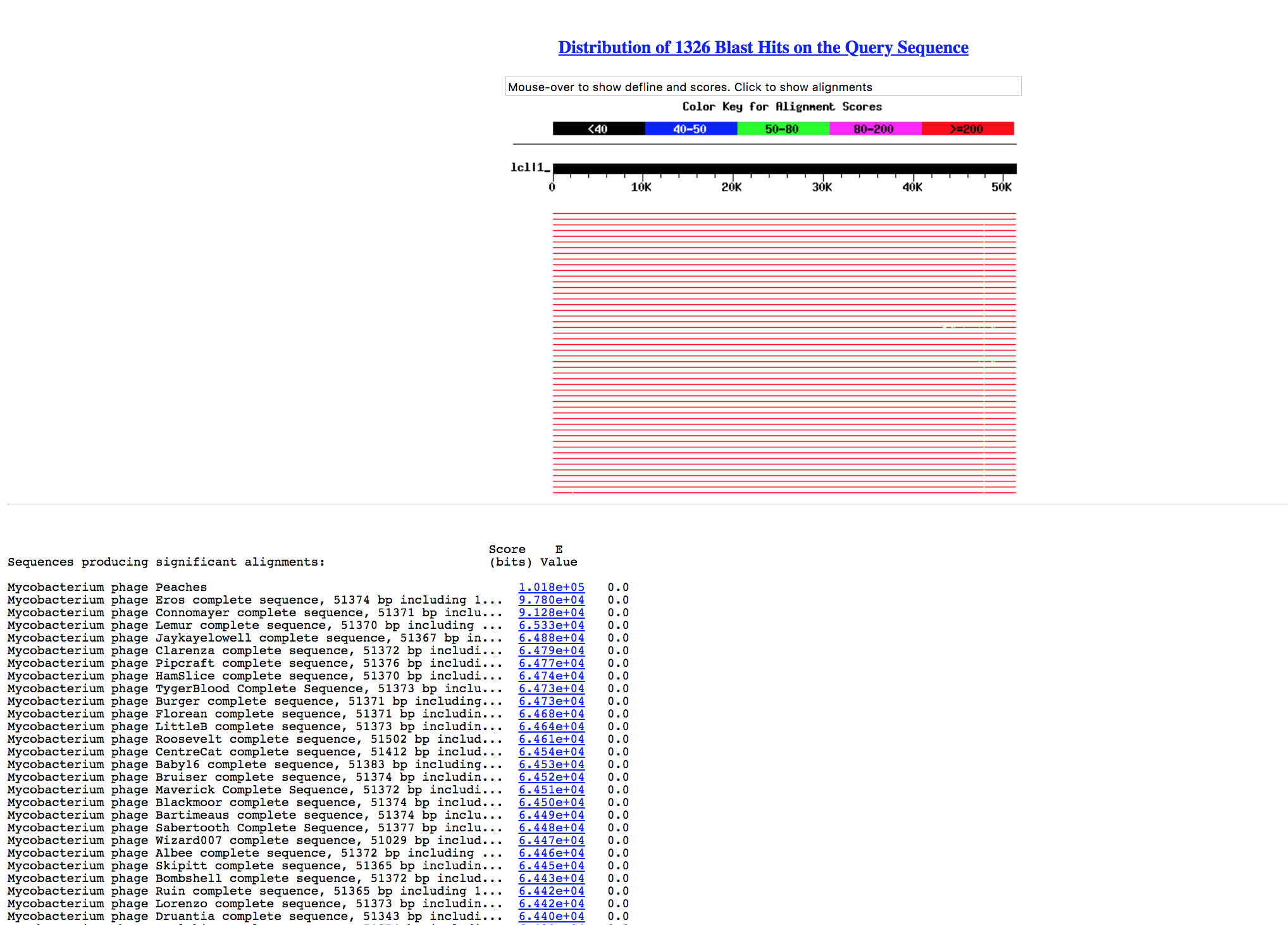
b. required adjustment of start or gene prediction

c. not informative

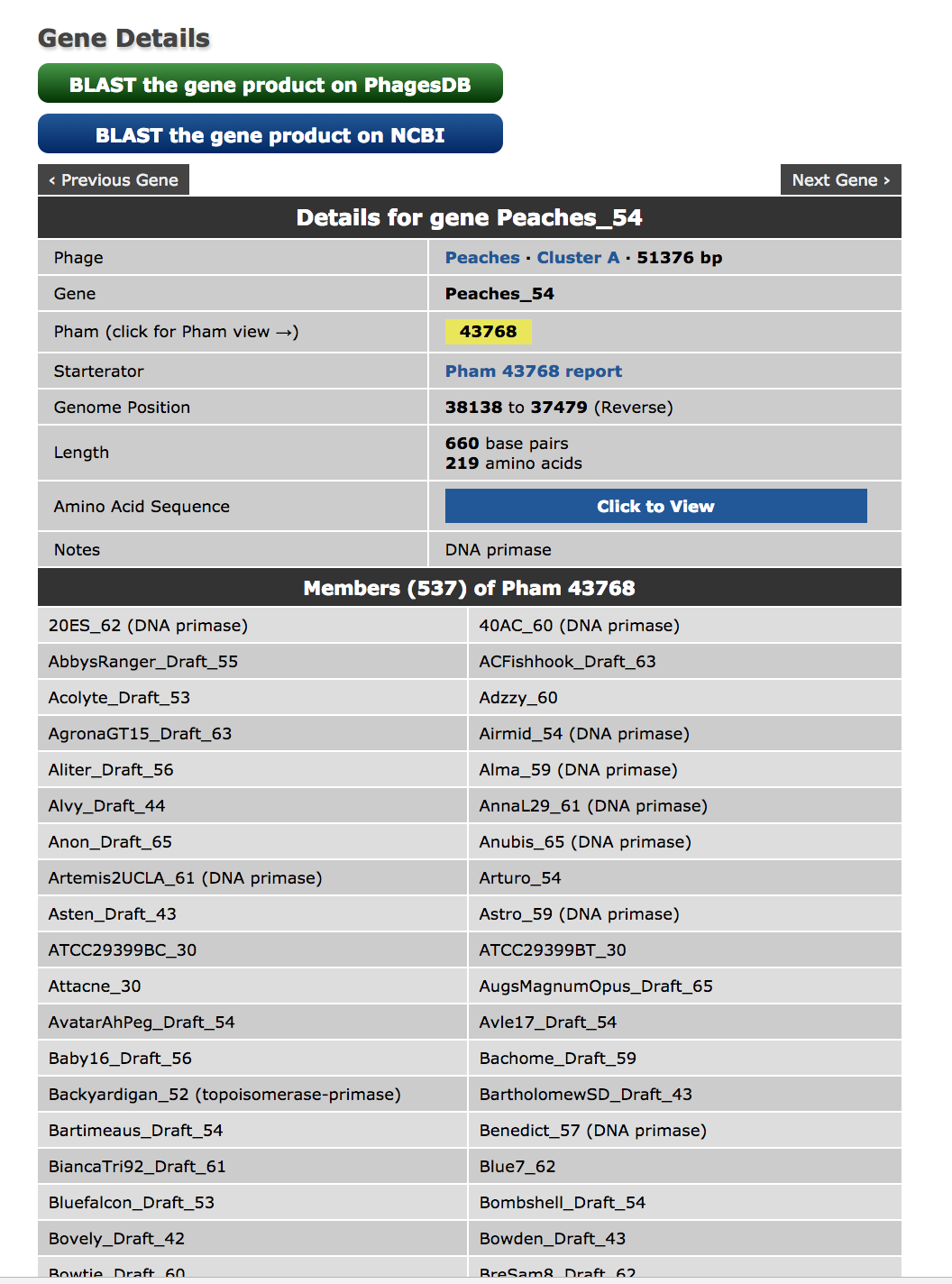
6. Does this function make sense in the context of the genome as a whole? Refer to the list of common phage functions, their most common genome locations, and the number of times they usually appear in a genome.

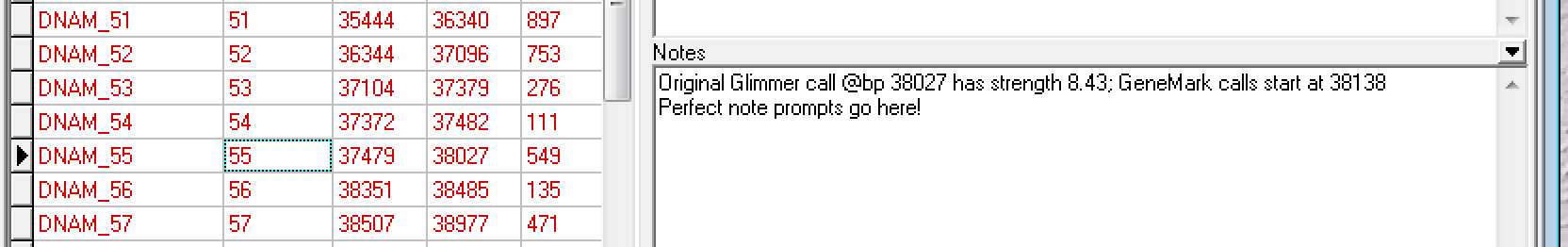
Yes, many phages have a DNA primase. It is usually only in one piece. However: The split DNA primase with a large overlap appears in every final version of a cluster A phage, and this is noted on the Cluster specific annotation forum.

A.

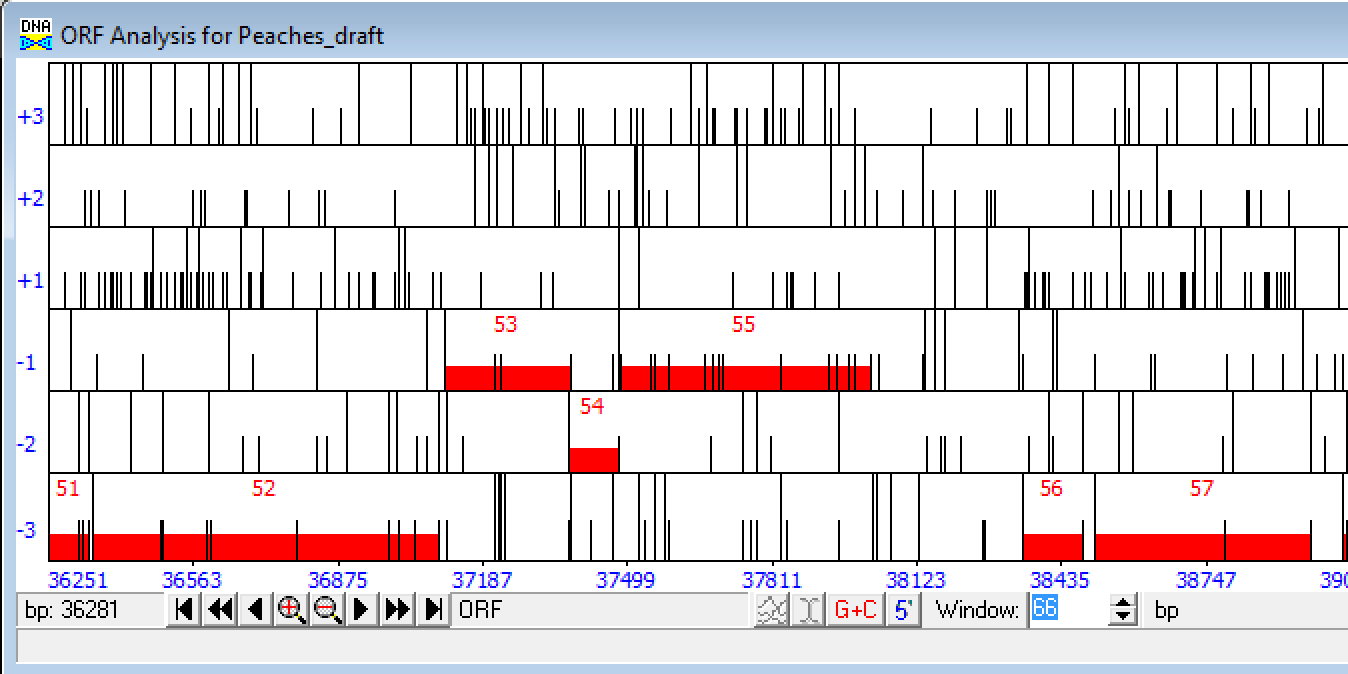


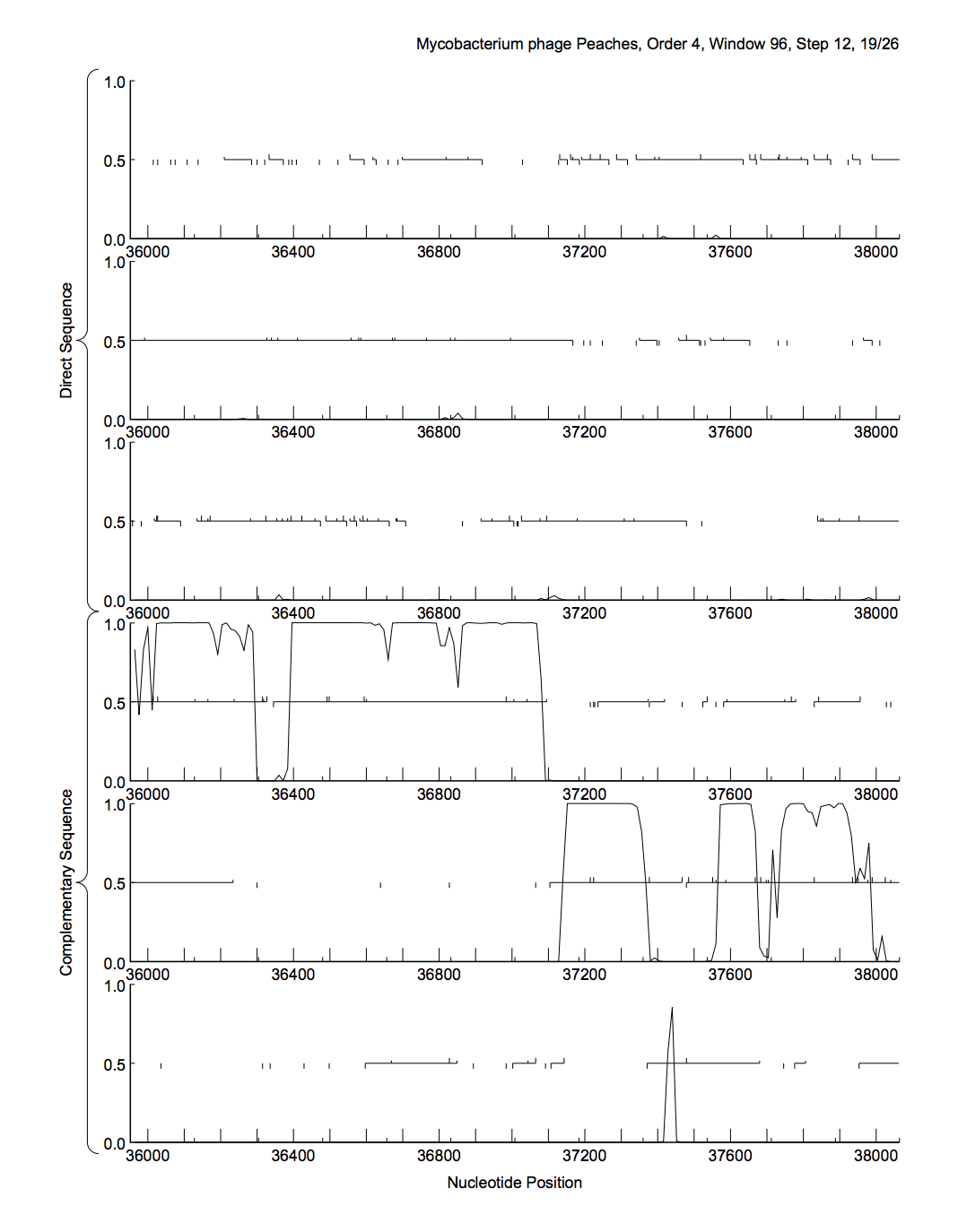
B.

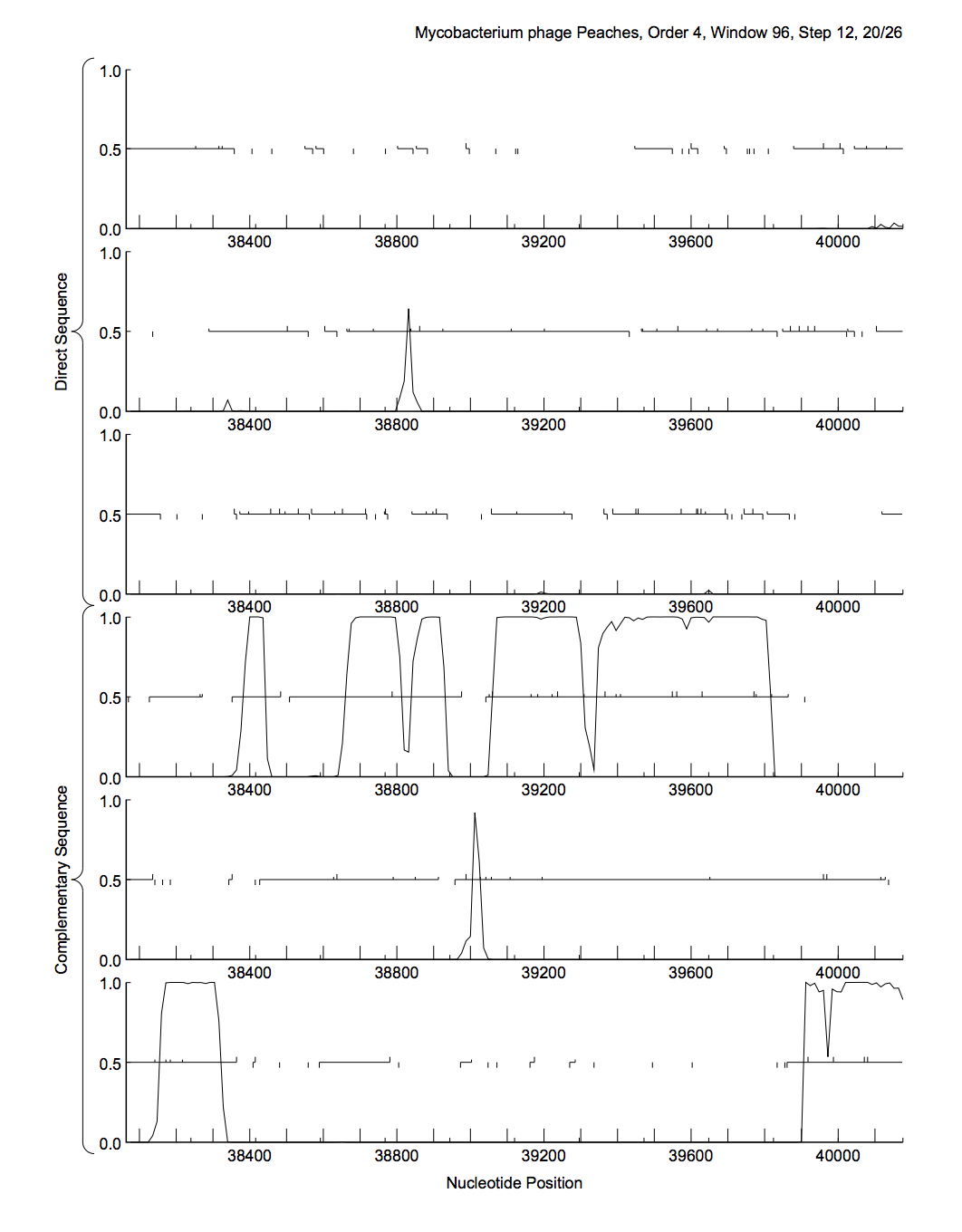


C. 

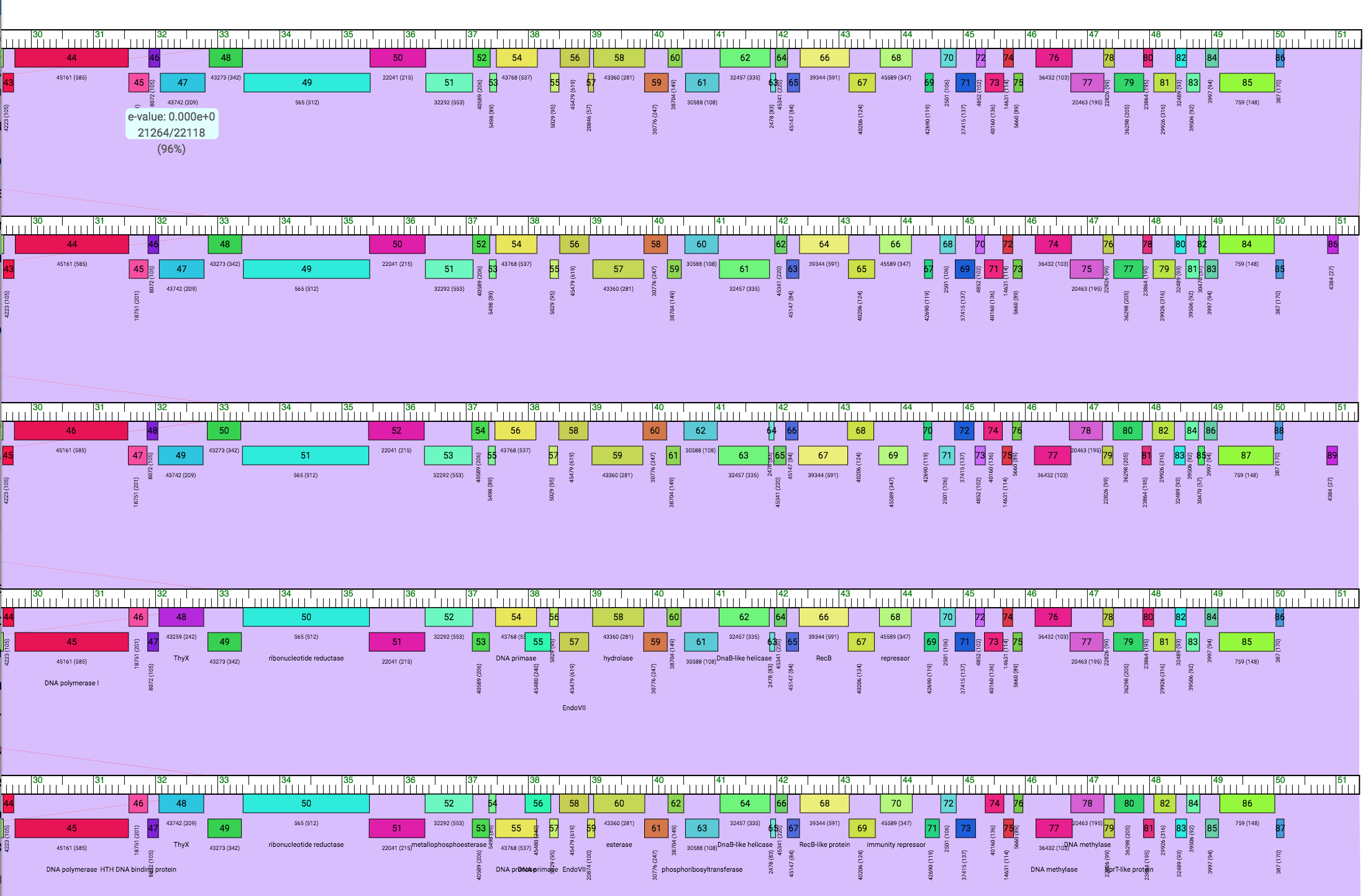
D.

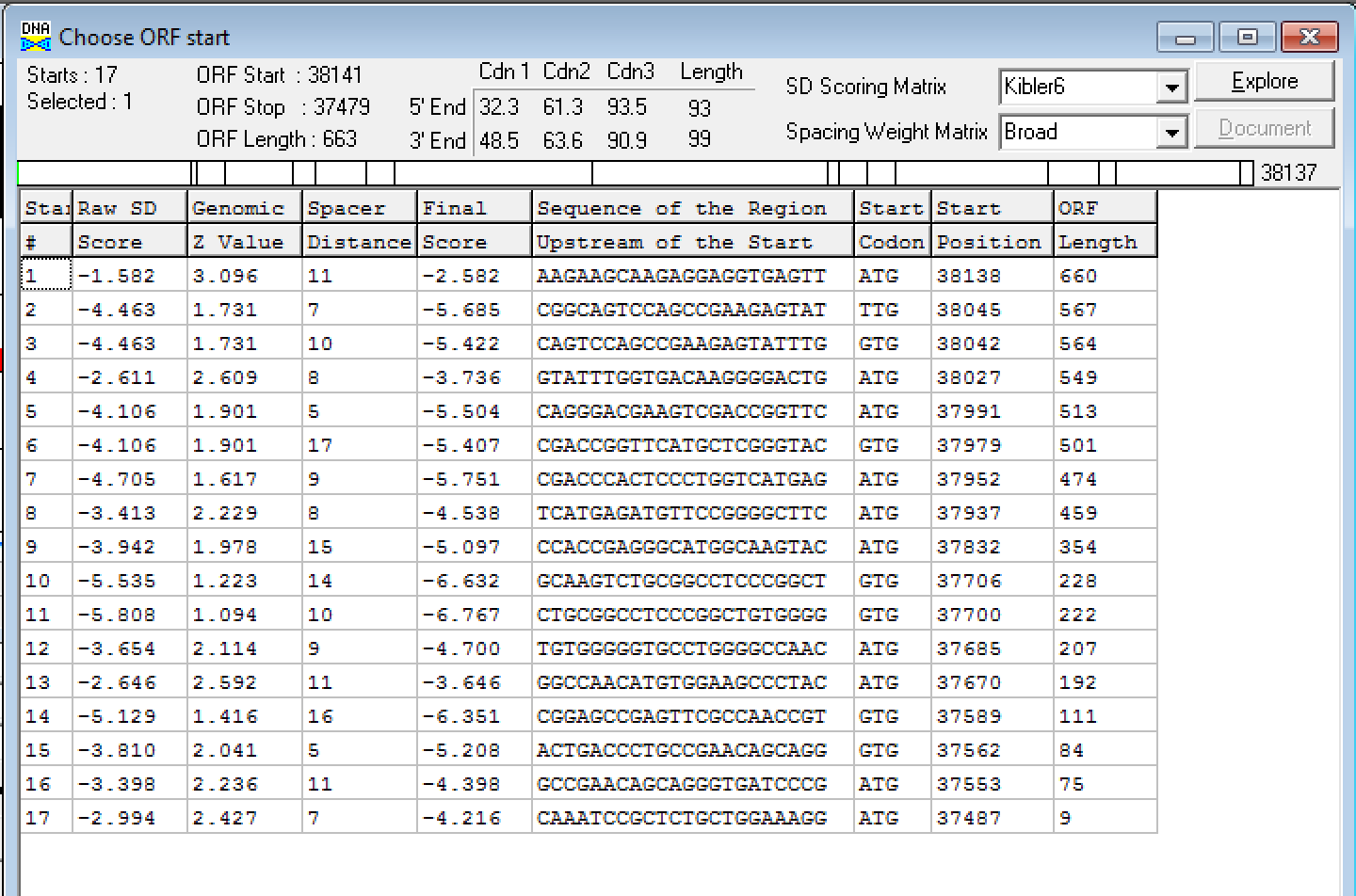


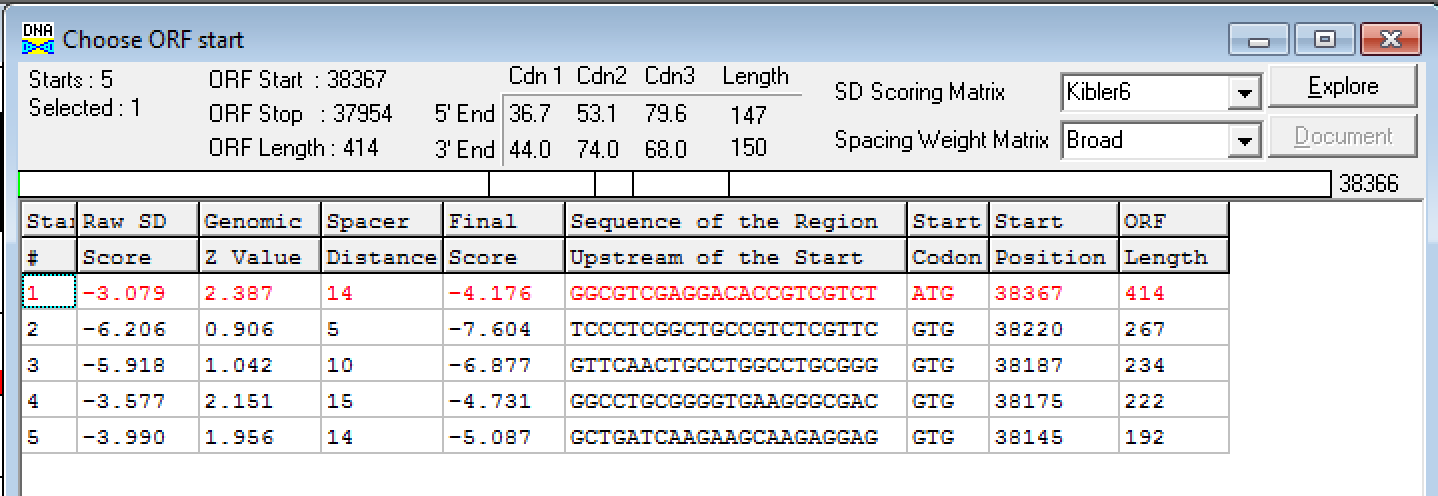
E.



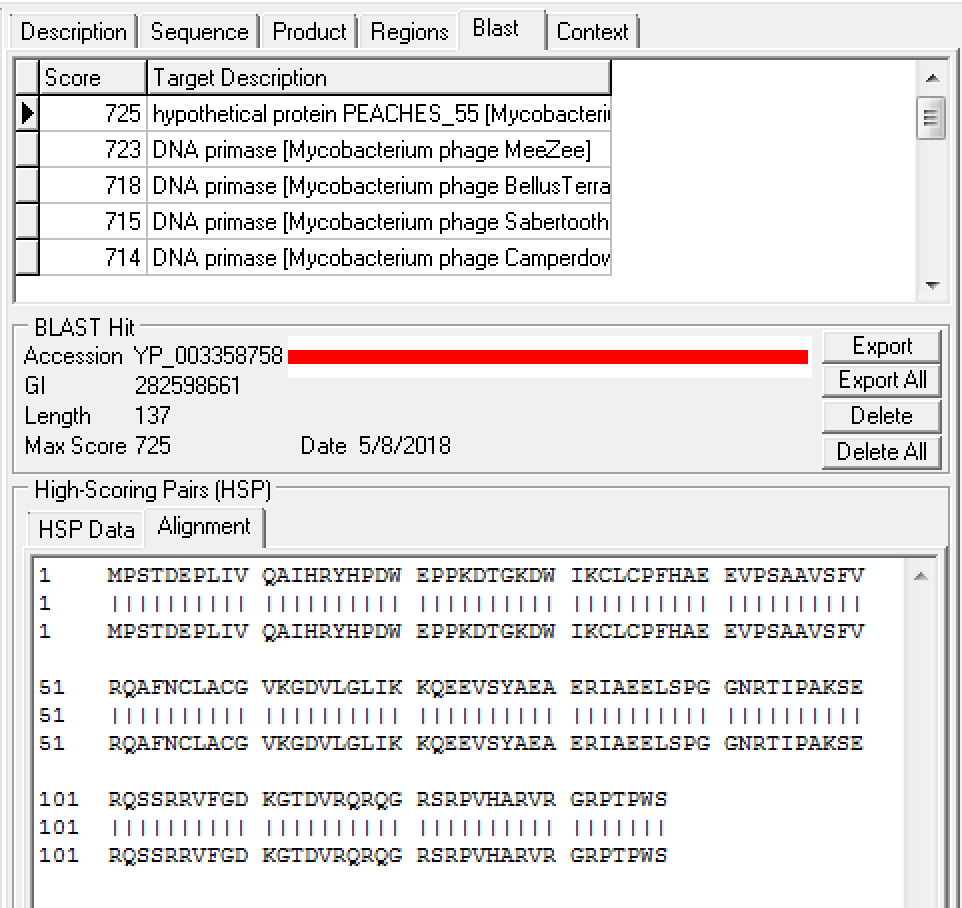
F:

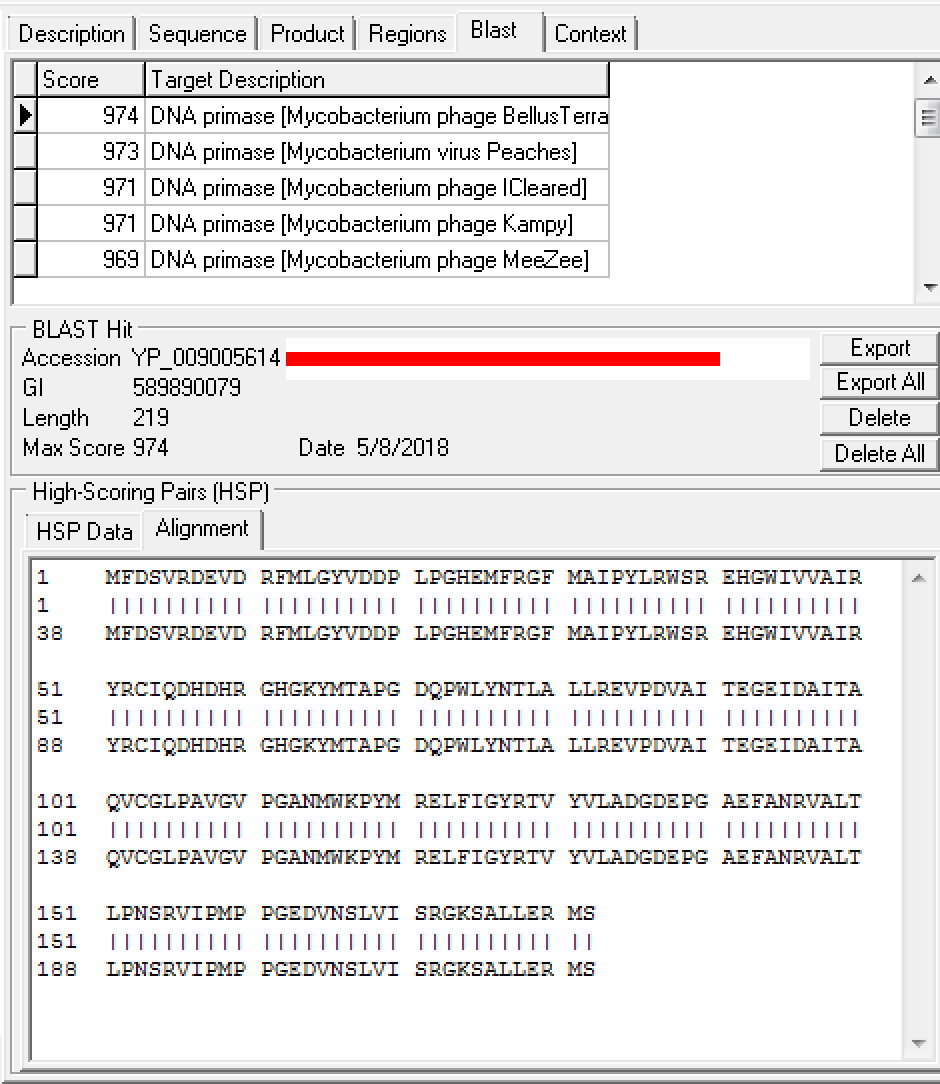


G: 



H.





I.



J.

