THOR cover sheet

In general we were rather conservative in changing things. No features have been added, and features were only deleted if there was no BLAST support and/or the feature was heavily overlapping a well-supported feature.

The areas that we think deserve closer attention are as follows:

1. The frame shift upstream of the tapemeasure. This is patterned almost directly on the SARFIRE frame shift (Sarfire appears to be the single most closely related virus). In SARFIRE, the two upstream genes have the exact same start/stop coordinates (i.e. overall) as in THOR, and in examining the relevant nucleotide sequences it definitely looked as if the shift would happen in exactly the same place. Still, it would be good to take a look at.

2. There are large gaps in the right hand end. None of these contain noteworthy GM coding potential, and in doing NCBI BLASTx only one gap reveals significant products in terms of prob. or e-values. Another reason they are being left alone is that doing otherwise seems to violate one of the main organizing principals of this genome: this organism’s genome clearly has a *leftward transcribing* (the right) and a *rightward transcribing* (the left) end. If you examine the gaps in the right end—the end where ABSOLUTELY EVERYTHING is transcribed going left—you see that there isn’t much opportunity to add something in the negative reading frames because they have abundant stop codons. The one decent looking place in the negative reading frames for doing so is the large gap just downstream from feature 700. However, this returns nothing meaningful (as noted above) when one does NCBI-BLASTx on the relevant sequence. The one large gap that did return good BLASTX data was in the vicinity of 37800; as above, the only reasonable place to add an ORF appears to be in a positive reading frame, violating as noted the overall organizing principal. Still you might want to take a look at it. BLASTx returned a hit to Sarfire in this region at high prob. and low E, but actually if you look at the Sarfire genbank file it looks like the only thing called there was in a negative frame, and it appears that nothing will fit there in THOR. Also please note that the far left 600bp has no coding potential; the far left end also has a 310bp gap after feature 30 with no called gene (and no coding potential). This structure for the far left end is consistent with closely related published Sarfire. Extreme right of genome likewise has no coding potential, and having no called gene at the far right is also consistent with Sarfire, which lacks any gene call in the extreme right kilobase.

In addition:

1. You may wish to take a look at features 560 and 910. Both are tiny (<100bo) but were left in

because they BLAST rather well and don’t overlap anything

2. An attempt to extend the ORF for feature 890 to the next available start actually resulted in

*poorer* BLAST results (see notes) so the call was returned to agreement with GM and GLM

3. The function of feature 600’s product was left NKF largely because BLAST was vague (see

notes)

4. The gaps flanking feature 360 are on the largish side (120 and 110bp) but there doesn’t seem

to be much that can be done about them.

5. Function call on feat. 650 is largely based on HHPred (only one BLAST hit)

6: Largish overlaps: 8bp overlap betw 310-320, 14bp overlap betw550-560, 8overlap betw580-

590; however it should be noted that overlaps seem to be consistent, in keeping with the above,

with the overall layout of this end of the genome.