Supplementary materials

Annotation revisions

Overview

Mass spectrometry data confirm the expression of 79 of the originally annotated genes in Patience, as well as one gene that is clearly expression but was not previously annotated. For 55 of the annotated genes, the chosen translational start site was confirmed. For 14 genes the mass spectrometry data did not provide information to support or revise the annotated start site. For seven proteins the translational start site was revised from the initial annotation. We also note that for the 46 instances in which the spectrometry data were informative, the N-terminal methionine was removed in 32 (leaving the residue encoded by the second codon at the extreme N-terminus, and for 14 the methionine remained at the N–terminus.

For those genes in which the translational start site could be determined, 46 use ATG, 12 use GTG and 4 use TTG.

Details are provided below.

New genes

A new gene (coordinates: 6881 – 6994; designated *111*) located between genes *8* and *9* was identified, with a total of 33 peptides.

Translation start site revisions

For gene *4* (new coordinates: 2042-2350) mass spec data support the use of an ATG codon at 2042 rather than at 2039. Many peptides were identified in which the alanine residue at the second codon of the newly annotated gene is at the N-terminus.

For gene *17* (new coordinates: 10120-10470) mass spec data support the use of an ATG start site at 10120 instead of at 10138. There are two peptides (at 70% and 90% confidence respectively) that support translation upstream of 10138. The new start site is accompanied by a strong ribosome binding site.

For gene *29* (new coordinates: 17715-18134) mass spec data support the use of the ATG codon at 17715 rather than the start at 17748. Many peptides were identified in which the alanine at the second position of the newly annotated gene is at the N-terminus.

For gene *47* (new coordinates: 38753- 39469), the mass spec data of early-infected cells supports translation initiation at the ATG codon at position 38753 instead of 38597. One peptide identified with 100% confidence remains at the N-terminus. The preceding residue (threonine) is not a site for tryptic cleavage. Heuristic GeneMark data support this call, but the TB GeneMark does not. A strong ribosome binding site lies upstream of the newly assigned start site.

For gene *53* (new coordinates: 41121-41441), the mass spec data of early-infected cells supports translation initiation at the ATG codon at position 41121, instead of at 41106. One peptide identified with 100% confidence remains at the N-terminus. The preceding residue (serine) is not a site for tryptic cleavage. Heuristic GeneMark data support this call, but the TB GeneMark does not. This is not predicted by GeneMark, but a strong ribosome binding site lies upstream of the newly assigned start site.

For gene *89* (new coordinates: 57623- 57922) the mass spectrometry data supports translation initiation at the ATG codon at position 57623 instead of 57680. Three peptides with probability scores of 100% are identified, all of which have a methionine at their N-terminus. The preceding residue (glutamine) is not a site for tryptic cleavage. Heuristic but not the TB trained GeneMark data support the use of this start site, and there is a good ribosome binding site upstream. Note that gene *89* now has a 62 bp overlap with the upstream gene, *88*.

For gene *101* (new coordinates: 64577-64987) mass spectrometry supports translation initiation at the ATG codon at 64577 instead of at 64667. Several peptides were identified corresponding to translation of the region upstream of 64667. The only upstream plausible start site is at 64577, although N-terminal peptides were not identified. The GeneMark coding potentials support this upstream start site and it is accompanied by a strong ribosome binding site. Note that gene 101 now overlaps by the upstream gene (*100*) by 64 bp.

Protein processing

For gp23 (capsid) no peptides were identified upstream of residue 83, consistent with proteolytic processing between residues 82 and 83. However, only a single peptide with residue 83 at its N-terminus was identified. The residue immediately upstream (proline) is not a site for tryptic cleavage.

For gp21, 11 peptides were identified (100% confidence) that are the most N-terminal, and contain a threonine residue corresponding to position 56 at their N-termini, which is acetylated. The preceding residue (glycine) is not a site for tryptic cleavage. A twelfth peptide is similar, but contains that glycine at the N-terminus. These suggest that the protein is proteolytically processed (albeit imperfectly) such that residue *56* is at the N-terminus of the mature protein.

Acetylation

For the following proteins the majority of N-terminal peptides are acetylated at their N-terminus: gp1, gp20, gp21, gp26, gp31, gp37, and gp79. Six of these are acetylated at a threonine residue, and one (gp37) at a serine.

Frameshift

Peptides encoded by gene *33*, the second open reading frame of the pair of tail assembly chaperones predicted to be expressed via a programmed translational frameshift, were identified. This includes peptides upstream of the annotated start site. However, the actual position of the +1 frameshift has not been identified.

Supplementary Files

The file Patience_ms-data.sf3 can be opened in the free Scaffold4 Viewer. It is available from: https://www.dropbox.com/s/s4r606g1vc5ybv0/Patience-ms-data.sf3?dl=0