SOP for Prediction of Influenza Protein Variants

1. Purpose

This document describes computational and scientific methods used to predict six influenza A virus (IAV) protein variants from genome sequence data:

a. PB1-N40 (Segment 2)
b. PA-N155 (Segment 3)
c. PA-N182 (Segment 3)
d. PA-X (Segment 3)
e. M42 (Segment 7)
f. NS3 (Segment 8)

Note that PB1-F2 (Segment 2) is provided by GenBank, and is not computed using any in-house method.

These influenza variant proteins can be retrieved from IRD for inspection at:

IRD Home > Gene/Protein Search > Protein Sequence Search > "Variant" Proteins

2. Method Descriptions

Methods for the prediction of each variant protein are described below. The type of variant is specified, along with the Reference Strain and Reference Accession used as a template model by the particular bioinformatics routine. The methods are embedded into the IRD data production environment and are not currently available as stand-alone scripts.

a. PB1-N40 (Segment 2)

For the PB1-N40 variant, an alternative start site is used, which is downstream of the first ATG start codon that initiates translation of the classical PB1 protein. Therefore the first 40 N-terminus amino acids of the classical PB1 are not translated in PB1-N40. In this prediction method, the alternative ATG start site is searched for at the correct position in the nucleotide sequence and, if detected, used to initiate translation, rather than simply deleting off the N-terminal amino acids of the native parent PB1 protein. There is no evidence that alternative splicing plays a role in the production of this protein. Only Influenza Type A is processed with this method.

Variation type:	Alternate Start
Reference Strain:	A/California/04/2009
Reference Accession:	GenBank accession JF915189

1. For each Segment 2 (PB1)
2. Align against A/California/04/2009
3. At position aligning with position 118 in A/California/04/2009, find ATG… (If ATG is absent, skip the sequence)
4. Translate
5. Confirm stop @~2272-2274 (if downstream stop doesn't exist, mark as "no stop codon detected" in database)
6. Confirm alignment with normal start PB1 for same frame

PB1-N40 was first described in reference 6: Wise *et al*. (2009) A complicated message: Identification of a novel PB1-related protein translated from influenza A virus segment 2 mRNA. *J Virol.* Aug; 83(16):8021-31. (PMID: 19494001). The predictive method is based on these data.

b. PA-N155 (Segment 3)

For the PA-N155 variant, an alternate start site (the eleventh AUG) is used in protein production. As a result, the amino terminal 155 residues are not present, compared to the classical PB1 protein. As above, a valid start site is searched for in the nucleotide sequence, and if detected, used to initiate translation, rather than simply deleting off the N-terminal amino acids of the native parent PB1 protein. There is no evidence that alternative splicing plays a role in the production of this protein. Only Influenza Type A is processed with this method.

Variation type:	Alternate Start
Reference Strain:	A/California/04/2009
Reference Accession:	GenBank accession FJ966081

1. For each Segment 3,
2. Align against A/California/04/2009,
3. Identify ATG at position aligning with position ~463-465 in A/California/04/2009 (If ATG is absent, skip the sequence),
4. Translate,
5. Confirm stop @~2149-2951 (if downstream stop doesn't exist, mark as "no stop codon detected" in database),
6. Confirm alignment vs. normal PA.

The reference model for the prediction of PA-N155 is built out of the results of Muramoto et al. (Identification of novel influenza A virus proteins translated from PA mRNA. *J Virol. 2013*) (5).

c. PA-N182 (Segment 3)

For the PA-N182 variant, an alternate start site (the thirteenth AUG) is used in protein production. As a result, the amino terminal 182 residues are not present, compared to the classical PB1 protein. As above, the start site is searched for in the nucleotide sequence, and if detected used to initiate translation rather than simply deleting off the N-terminal amino acids of the native parent. There is no evidence that alternative splicing plays a role in the production of this protein. Only Influenza Type A is processed with this method.

Variation type:	Alternate Start
Reference Strain:	A/California/04/2009
Reference Accession:	GenBank accession FJ966081

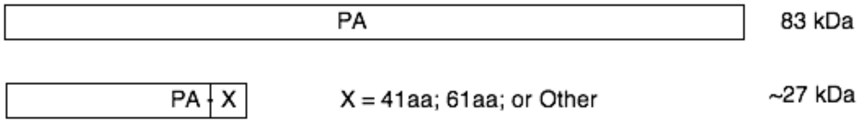
1. For each Segment 3,
2. Align against A/California/04/2009,
3. Identify ATG at position aligning with position ~544-546 in A/California/04/2009 (if ATG exists, continue; if ATG is absent, skip the sequence),
4. Translate,

5. Confirm stop @2149-2151 (if downstream stop doesn't exist, mark as "no stop codon detected" in database)
6. Confirm alignment vs normal PA.

As for PA-155, the reference model for the prediction of PA-N182 is built out of the results of Muramoto et al. (Identification of novel influenza A virus proteins translated from PA mRNA. *J Virol. 2013*) (5).

d. PA-X (Segment 3)

The PA-X protein is produced as a direct result of a ribosomal stutter and frameshift to a new +1 reading frame, resulting in a shorter protein overall, with altered sequence downstream, and altered structure and function. The role of the 27 kDa PA-X protein and its variants in viral replication are not fully understood, but are thought to modulate IAV pathogenicity to the host. The PA variants are depicted below, with the full length 83 kDa PA on top, and the variants on bottom:



The length of the "X" extension is determined by the nucleotide sequence, which is translated until a stop codon is encountered in the new reading frame. Therefore, a more complex computational method was developed to predict the presence of PA-X protein sequences, whereby the ribosomal stutter site is recognized, the appropriate frameshift made, and the variant protein translated, as detailed below:

Variation type:	Alternate Splice
Reference Strain:	A/WSN/1933
Reference Accession:	GenBank CY034137

1. For each Segment 3,
2. Perform global alignment against Segment 3 from reference strain A/WSN/1933,
3. PA nucleotide position numbers are taken from the A/WSN/1933 strain (GenBank CY034137),
4. The 'slippery' or stutter motif '**TCC-TTT-CGTC**' near the C598 residue is identified,
5. A yes/no determination is made to proceed. The candidate motif must meet three conditions:
    i) The sequence must possess a TTT Phe codon at aligned position 595-597,
    ii) TTT is followed immediately by a C, resulting in a Phe 'wobble' codon of TTC in the new ribosomal stutter frame (+1). TTG and TTA are not allowed (Leu).
    iii) the new X amino acid must translate as Val, the only instance known (2).
7. Translate,
8. Annotate as +41, +61, other;
9. Confirm stop.

The above conditions were established based on following biology. The TTT codon (red below, underlined) of the conserved sequence motif encodes the Phe amino acid. During PA translation, the host ribosome stalls at that TTT codon while waiting for the next "rare" codon CGT to be filled with a charged tRNA. While waiting, the ribosome 'wobbles' forward to the alternate Phe codon TTC or TTT in approximately 1-2% of all translations, producing a +1 frameshift event. C598 is defined as producing PA-X, whereas A598 and G598 are disallowed. At the time of writing, no T598 has been identified in the database or literature, and while the stutter forward would produce a Phe at that position, the method throws out the possibility of PA-X expression since that resulting protein has not yet been experimentally shown to exist.

In summary Phe is ***still*** encoded (underlined, black) after the stutter but now in a new frame, as below:

```
Segment 3 Nucleotide: TCCTTTCGTC
Amino acid (PA):      SerPheArg…
Amino acid (PA-X):    SerPhe-Val…
```

The dash in the third line represents the 'skipped' cytosine C after TTT, with the next GTC encoding a Val in the "X" extension, instead of the native Arg encoded by CGT. Furthermore, the sequence is then translated in the new frame until a stop codon is reached. If both an authentic start and stop codon are **not** found, the sequence is termed ambiguous, the protein not predicted, and therefore not displayed in IRD as a new PA-X protein. Only Influenza Type A is processed in this method.

e. M42 (Segment 7)

In this rare variant (~0.1%), both an alternate start codon and alternative splice donor and acceptor sites are used. Only Influenza Type A is processed in this method.

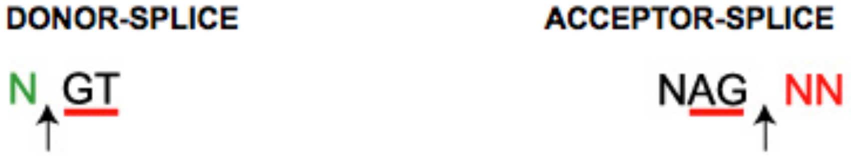
Variation type:	Alternate Start; then Splice
Parent Segment:	7 (M1/M2)
Reference Strain:	A/Puerto Rico/8/34 strain
Reference Accession:	GenBank accession EF467824

1. Verify as Segment 7 (M1M2),
2. Align vs. (A/Puerto Rico/8/34),
3. Identify ATG start site at position aligning with position ~114-116 in A/Puerto Rico/8/34 (…GAAG^ATG…),
4. Translate forward in this frame through position 145 (…ILR^),
5. Skip (splice out) intron up to position 740 (^PIR…) ,
6. Translate,
7. Confirm stop @ position ~1005-1007 (if downstream stop doesn't exist, mark as "no stop codon detected" in database),
8. Confirm second exon aligns to M2 translation, in the same frame.

Because the M2 and M42 proteins share the 3' exon, the protein sequence that is thereby predicted for these two proteins in this region should be identical. It is only the N-terminus of the protein that differs between M2 and M42. M42 was first described in reference 7: Wise *et al.* "Identification of a Novel Splice Variant Form of the Influenza A Virus M2 Ion Channel with an Antigenically Distinct Ectodomain", *PLoS Pathogens* 8(11), 2012. (PMID: 23133386).

f. NS3 (Segment 8)

For this exceedingly rare variant (~0.05%), a random sequence polymorphism introduces a novel splice site into the NS1 gene. The resulting protein produced from the alternatively spliced transcript is called NS3. The prediction method searches for a valid splice site near codon 124 of NS1, by checking against donor and acceptor rules, where the splice donor must END in ^GT, and the splice acceptor must START in AG^, as follows:



Variation type: Alternate Splice
Reference Strain: A/Udorn/8/1972
Reference Accession: GenBank accession V01102

Where the method is as follows:

1. For each Segment 8 (NS),
2. Align to A/Udorn/8/1972 NS1,
3. Translate from start codon,
4. At position aligning with position ~373 (codon 124) in A/Udorn/8/1972, identify presence of sequence polymorphism that determines splicing or not,
5. If variant GGT (glycine/G), Translate G and pause (this is a new splice donor site) Goto #7, else,
6. If native GAT (aspartate/D) is present, stop process, throw out sequence, and exit,
7. Move forward to the splice acceptor site position aligning with position ~503 in A/Udorn/8/1972 (amino acid 169),
8. Begin in frame translate the second exon,
9. Confirm stop (if downstream stop doesn't exist, mark as "no stop codon detected" in database),
10. Confirm alignment to NS1.

This method is based on the presence of rare polymorphisms that produce a splice motif (7). Only Influenza Type A is processed with this method.

3. Input Data Preparation

Input data is in the form of GenBank files, downloaded nightly. These methods takes these files as input, as is. Bioinformatics scripts are run on SQL-based extracts from the IRD database, which have passed influenza autocuration QA checks.

4. Output Data, Processing, and Display

New sequences are entered in the database by the autocuration routines. Processed sequences are labeled according to whether they are predicted to produce variant proteins. In the case of PA-X, the length of the X is also noted as +41, +61, or other. A menu to search for and display variant proteins is made available in the IRD "protein search" interface.

5. References

1. Jagger, B.W., Wise, H.M., Kash, J.C., Walters, K.A., Wills, N.M., Xiao, Y.L., Dunfee, R.L., Schwartzman, L.M., Ozinsky, A., Bell, G.L., Dalton, R.M., Lo, A., Efstathiou, S., Atkins, J.F., Firth, A.E., Taubenberger, J.K., Digard, P. (2012) An overlapping protein-coding region in influenza A virus segment 3 modulates the

host response. Science Jul 13;337(6091):199-204. Epub Jun 28. PubMed PMID: 22745253.

2. Yewdell, J.W., Ince, W.L. (2012) Virology: Frameshifting to PA-X influenza. Science Jul 13;337(6091):164. PMID: 22798590.

3. Bakeart, M. and Viralzone. (2012) "Viral Ribosomal Frameshifting" URL: http://viralzone.expasy.org/all_by_protein/860.html

4. Firth A.E., Brierley, I. (2012) Non-canonical translation in RNA viruses. J Gen Virol. Jul;93(Pt 7):1385-409. Epub Apr 25. PubMed PMID: 22535777

5. Muramoto Y, Noda T, Kawakami E, Akkina R, Kawaoka Y. Identification of novel influenza A virus proteins translated from PA mRNA. J Virol. 2013 Mar;87(5):2455-62. PMID: 23236060.

6. Wise *et al.* A complicated message: Identification of a novel PB1-related protein translated from influenza A virus segment 2 mRNA. J Virol. 2009 Aug; 83(16):8021-31. (PMID: 19494001).

7. Wise *et al.* "Identification of a Novel Splice Variant Form of the Influenza A Virus M2 Ion Channel with an Antigenically Distinct Ectodomain", *PLoS Pathogens* 8(11), 2012. (PMID: 23133386).