SOP for Influenza autocuration

Authors: Catherine Macken (c.macken@auckland.ac.nz), Sam Zaremba (Sam.Zaremba@ngc.com), Guangyu Sun (gsun@vecna.com), Sherry He (she@virusbrc.org), Christian Suloway (csuloway@virusbrc.org), Dennis Woodbine (dwoodbine@virusbrc.org), David Suarez (David.Suarez@ARS.USDA.GOV), Richard Scheuermann (RScheuermann@jcvi.org)

Submission date: July 31 2017
Version: 2.0

I. Overview
IRD uses an automated pipeline to ascertain the presence of potential sequence artifacts or other possible contributors to poor quality in Influenza virus nucleotide sequences. The basis for curation of a sequence of a given type/segment/subtype is its alignment against a reference alignment of sequences from the same type/segment/subtype. The reference alignments are carefully developed to include the range of legitimate variation for the given type/segment/subtype collection of sequences. Every sequence in IRD undergoes this quality control during the loading process. The pipeline sets flags indicating the category and location of potential sequence artifacts, or the type of poor quality sequence. These flags are summarized then displayed to an IRD user in sequence search results and in listings of working set contents (see Section II.3); details of the summary flags are shown on Segment/Protein Details pages. The pipeline is also invoked when data providers prepare and submit sequences to Genbank by using the IRD web site. Any flags raised by the submitted sequence(s) are reported to the submitter, who can then use them to correct sequences before final submission.

After curating a sequence, its aligned version is stored in the database. IRD users can then use these pre-computed aligned sequences in their downstream analysis (MSA, SNP, meta-CAT, phylogenetic tree) and benefit from high quality of the pre-computed aligned sequences and reduced compute times, especially for large datasets.

II. Process

1. Profile alignment with Clustal W
To assess the quality of an influenza nucleotide sequence (here called a “query”), a preliminary analysis uses BLAST to identify the type and segment (and subtype for Influenza A HA segment 4 and NA segment 6) of the query. The query is then aligned, using the profile-alignment protocol of Clustal W, against a curated reference alignment (here called a “profile”) that captures the natural variation in the appropriate type/segment/subtype category. These profiles are annotated by nucleotide coordinates for 5’ and 3’ Conserved Terminal Sequences (CTS) [Desselberger et al, 1980], 5’ and 3’ Non-Coding Regions (NCR), and Coding Sequence (CDS), allowing IRD to locate potential artifacts within each of the respective sequence regions. The robustness of these profiles is monitored regularly, and profiles are updated by the IRD staff when new sequence patterns emerge and are validated in multiple sequences and sequencing laboratories.

The alignment of a query to its appropriate profile, and assignment of flags, are illustrated for a hypothetical profile and query sequence in Figure 1.
**Figure 1a: 5’ end of alignment of query to profile**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Query</th>
<th>AB212052_p</th>
<th>CY003990_p</th>
<th>KCR66610_p</th>
<th>CY100446_p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TTAT</td>
<td>AGCG</td>
<td>AGCG</td>
<td>AGCG</td>
<td>AGCG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AAAG</td>
<td>AAAG</td>
<td>AAAG</td>
<td>AAAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCAG</td>
<td>GCAG</td>
<td>GCAG</td>
<td>GCAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CAGC</td>
<td>CAGC</td>
<td>CAGC</td>
<td>CAGC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Profile sequences are indicated by “_p” after an accession. Regions of the profile alignment are indicated below, and flags for potential sequence artifacts are indicated above, the sequences. The 5’CTS domain is conserved with the possible exception of position 4. The 5’NCR beyond the CTS has limited variation. The CDS starts with a conserved start codon (ATG). The hypothetical query sequence has a 5’ extension of 3 bases, which are not from an influenza virus. Other potential sequence artifacts are indicated.
Regions of the profile alignment (here only CDS) are indicated below, and flags for potential sequence artifacts are indicated above, the sequences. The hypothetical query sequence has potential sequence artifacts as indicated. Insertions or deletions of 1 or 2 bases cause a frame shift and are flagged. The presence of a novel 3-base deletion will trigger a review by IRD staff.

2. Artifact flags

Table I lists all the flags used by IRD to describe potential sequence artifacts, including those illustrated in Figures 1a and b. It also lists summary flags, which appear in listings of search results and contents of working sets. IRD leverages the outcome of the curation process as described by these flags to offer new choices to users for deciding whether and how to include problematic sequences in their analyses (Section III).

Table I

<table>
<thead>
<tr>
<th>Region</th>
<th>Artifact Flag</th>
<th>Definition</th>
<th>Summary Flag</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTS</td>
<td>5’CTS-del</td>
<td>Deletion of any number of nt in 5’CTS</td>
<td>Flag-NCR</td>
</tr>
<tr>
<td>CTS</td>
<td>5’CTS-mut</td>
<td>Mutation of nt in 5’CTS</td>
<td>Flag-NCR</td>
</tr>
<tr>
<td>CTS</td>
<td>3’CTS-ins</td>
<td>Insertion of any number of nt in 3’CTS</td>
<td>Flag-NCR</td>
</tr>
<tr>
<td>CTS</td>
<td>3’CTS-del</td>
<td>Deletion of any number of nt in 3’CTS</td>
<td>Flag-NCR</td>
</tr>
<tr>
<td>CTS</td>
<td>3’CTS-mut</td>
<td>Mutation of nt in 3’CTS</td>
<td>Flag-NCR</td>
</tr>
<tr>
<td>NCR</td>
<td>5’NCR-ext</td>
<td>Extra nt preceding the 5’NCR, possibly from primer</td>
<td>Flag-NCR</td>
</tr>
<tr>
<td>NCR</td>
<td>3’NCR-ext</td>
<td>Extra nt trailing the 3’NCR, possibly from primer</td>
<td>Flag-NCR</td>
</tr>
<tr>
<td>NCR</td>
<td>5’NCR-ins</td>
<td>Insertion of any number of nt in 5’NCR, excluding 5’CTS</td>
<td>Flag-NCR</td>
</tr>
<tr>
<td>NCR</td>
<td>5’NCR-del</td>
<td>Deletion of any number of nt in 5’NCR, excluding 5’CTS</td>
<td>Flag-NCR</td>
</tr>
<tr>
<td>NCR</td>
<td>3’NCR-ins</td>
<td>Insertion of any number of nt in 3’NCR, excluding 3’CTS</td>
<td>Flag-NCR</td>
</tr>
</tbody>
</table>
Table II defines additional flags, which alert the user to problems with the query over its entire length.

### Table II

<table>
<thead>
<tr>
<th>Region</th>
<th>Artifact Flag</th>
<th>Definition</th>
<th>Summary Flag</th>
</tr>
</thead>
<tbody>
<tr>
<td>entire</td>
<td>HammDist</td>
<td>Insufficient match to profile sequences by Hamming Distance (Hamming, 1950)</td>
<td>Ambig-Seq</td>
</tr>
<tr>
<td>entire</td>
<td>Excess-N</td>
<td>More than 0.5% nucleotides in the sequence are N's</td>
<td>Ambig-Seq</td>
</tr>
<tr>
<td>entire</td>
<td>Excess-Ambig</td>
<td>More than 0.5% nucleotides in the sequence are not A,C,G,T,N</td>
<td>Ambig-Seq</td>
</tr>
</tbody>
</table>

3. **Logic and display of Summary flags**

The logic by which a sequence is assigned a Summary Flag is as follows:

1. If a sequence has one or more of the flags HammDist, Excess-N, or Excess-Ambig, it is assigned a Summary Flag = Ambig-Seq, regardless of whether or not it also has flags in non-coding regions or CDS.
2. If a sequence is not flagged Ambig-Seq, but has any flag in the CDS, it is assigned a Summary Flag = Flag-CDS, regardless of whether or not it also has flags in non-coding regions.
3. If a sequence is not flagged Ambig-Seq or Flag-CDS, but has any flags in non-coding regions, conserved terminal sequences, or extensions, it is assigned a Summary Flag = Flag-NCR.
4. Query sequences that align to the appropriate profile without triggering any of the above flags are assigned a Summary Flag = Pass.

IRD users will find Summary Flags accompanying each sequence record on Search Results pages and in listings of the contents of working sets (Figure 2).

**Figure 2: Listing of contents of working set showing curation flags**
4. Artifact details for a sequence record

Details of flags, including their position in the query sequence, are provided in the Curation Report on the IRD Sequence Details Page (Figure 3).

Figure 3: Details of flags on sequence details page

The position of each flag is counted from the beginning of the query sequence. Thus, a 3\'NCR-del flag at position 1363..1364 indicates 2 missing bases between positions 1363 and 1364 from the 5\’ end of the query sequence.
5. Pre-aligned sequences

The pre-aligned-to-profile sequence of each query, as produced by the curation pipeline, (Figure 4) is saved in the IRD database. In general, analyses that require aligned sequences (e.g., SNP analyses) are faster and more accurate when based on these IRD aligned sequences than when based on alignments obtained from the raw sequences using MUSCLE or similar programs. However, care must be taken when using sequences having a summary flag other than “pass” (Section III). Sequences with a summary flag “Flag-CDS”, indicating potential artifacts in the coding region, must be realigned by the user. Sequences with flags indicating potential artifacts in the CTS/NCR regions of a segment, but not in the CDS, have been edited by IRD, as described in Section II-6, to allow their use in alignment with only a small loss of information.

![Figure 4a: The 5’ end of an Influenza sequence as submitted to a public database. The initiation codon is in red.](image)

![Figure 4b: The same sequence after alignment in the IRD pipeline. The first 30 nts of the profile, the 5’CTS and 5’NCR, were absent from the submission and are indicated with ~. There are two 3X-deletions, which may or may not be novel.](image)

6. Edited Flag-NCR pre-aligned sequences

When the IRD pipeline assigns the summary “Flag-NCR” to a sequence, it also generates and saves an edited version of the aligned-to-profile sequence. These edited versions allow investigators to take advantage of the speed and quality of analyses using pre-aligned sequences (Section III) while including a higher proportion of their dataset in the analysis. Edited versions are constructed as follows:

- If a sequence only has 5’NCR-ext or 3’NCR-ext flags (see Table I), the superfluous preceding or trailing nucleotides are removed from the aligned-to-profile sequence. This leaves the aligned CTS, NCR and the CDS intact.

- If a sequence has one or more of the flags NCR-del, NCR-ins, CTS-del, CTS-ins and CTS-mut (see Table I), but no flags in the CDS and no Table II flags, the entire affected NCR domain(s) is(are) replaced by blanks in the edited pre-aligned version. In addition, any NCR-ext bases are trimmed. This leaves any unaffected NCR region, as well as the CDS, intact.

Edited versions are not generated for aligned-to-profile sequences having “Flag-CDS” or “Ambig-Seq” summaries.

III Leveraging the pipeline results for IRD analyses

When IRD sequences are used in any IRD analysis tool that relies on multiple sequence alignments (e.g. SNP, meta-CAT, phylogenetic tree) we recommend using the pre-aligned versions (possibly edited) of
the sequences of the input set. In extensive testing these pre-aligned sequences provided quality as good as, or substantially better than, those produced by running MUSCLE de novo on the same dataset of unaligned sequences. (The quality of an alignment is sensitive to the amount of internal gaps, so that, for example, a MUSCLE alignment of “gappy” A/N2 sequences can be quite poor.) Furthermore, by avoiding the time-consuming step of aligning using MUSCLE, analysis times were reduced from days to hours for large input sets. When a dataset from a search or personal workbench Working Set (WS) is selected for analysis, IRD offers users the following choices for their analyses (Figure 5).

- (Recommended 1) Use pre-aligned sequences. Include inputs with Summary Flag=Pass and the edited versions of any inputs with Summary Flag= Flag-NCR. Exclude all inputs with Summary Flag= Flag-CDS or Ambig-Seq.
- (Recommended 2) Use pre-aligned sequences. Include only those inputs with Summary Flag=Pass. Exclude all inputs with curation concerns, i.e., Summary Flag= Flag-NCR or Flag-CDS or Ambig-Seq.
- (Not recommended) Do not use pre-aligned sequences. Include everything in the input set and realign with MUSCLE.

Figure 5: SNP Analysis Setup page, with options for using (possibly edited) pre-aligned sequences
References:
