

1. Purpose

This document describes the mechanism and process by which polymorphism quantification data are computed for nucleotide sequences, and then stored in the IRD and ViPR database.

2. Method Description

The polymorphism data production pipeline is composed of two PERL scripts. All sequences were downloaded from GenBank. Influenza sequences were processed through the IRD curation pipeline. During this processing, pre-aligned sequences were generated with the ClustalW multiple alignment tool (1). For flu, aligned sequences from the same host and segment (and subtype) were used to determine sequence polymorphism. In order to identify polymorphisms with frequency of 0.1 or more, each multiple alignment was required to include at least 20 sequences. Because partial sequences were allowed, the number of sequences at some positions might be less than the number of sequences used for multiple alignments. In many cases, the positions towards the two ends had the least sequence coverage and as a result might appear to be more conserved than they actually are.

After the multi-alignment step, a **consensus sequence** for each group was created by following the "majority rule". For each position in the multiple alignments, the consensus was the allele with frequency greater than 50%, regardless of the sequence coverage. If a consensus could not be found (no allele had > 50% frequency), an N (for nucleotide) or Xaa (for amino acid) was used to indicate ambiguity. By comparing the consensus with each sequence in the alignment, we identified the nucleotide differences (SNP and indels) between them.

To quantify the nucleotide polymorphism at each position, a score was generated by using a formula modified from the one as described in Crooks et al (2), as follows:

$$S = -100 * \text{Sum} (P_i * \log_2 P_i)$$

where P_i is the frequency of the i th allele encountered.

The score S has a defined range. Basically, it is the normalized entropy of the observed allele distribution at each position. The **least** polymorphic site would have a score of **0** (single allele) and the **most** polymorphic site would have a score of **232** (e.g. 4 alleles with 20% frequency each, plus a deletion allele of the same frequency).

A position on a genome consensus sequence can be either coding or non-coding depending on whether it is part of the region that encodes any known protein product. We have chosen EXONERATE (<http://www.ebi.ac.uk/~guy/exonerate>) to align the consensus sequence with its cognate protein products. We then classify each position as coding or non-coding based on the DNA-protein alignment.

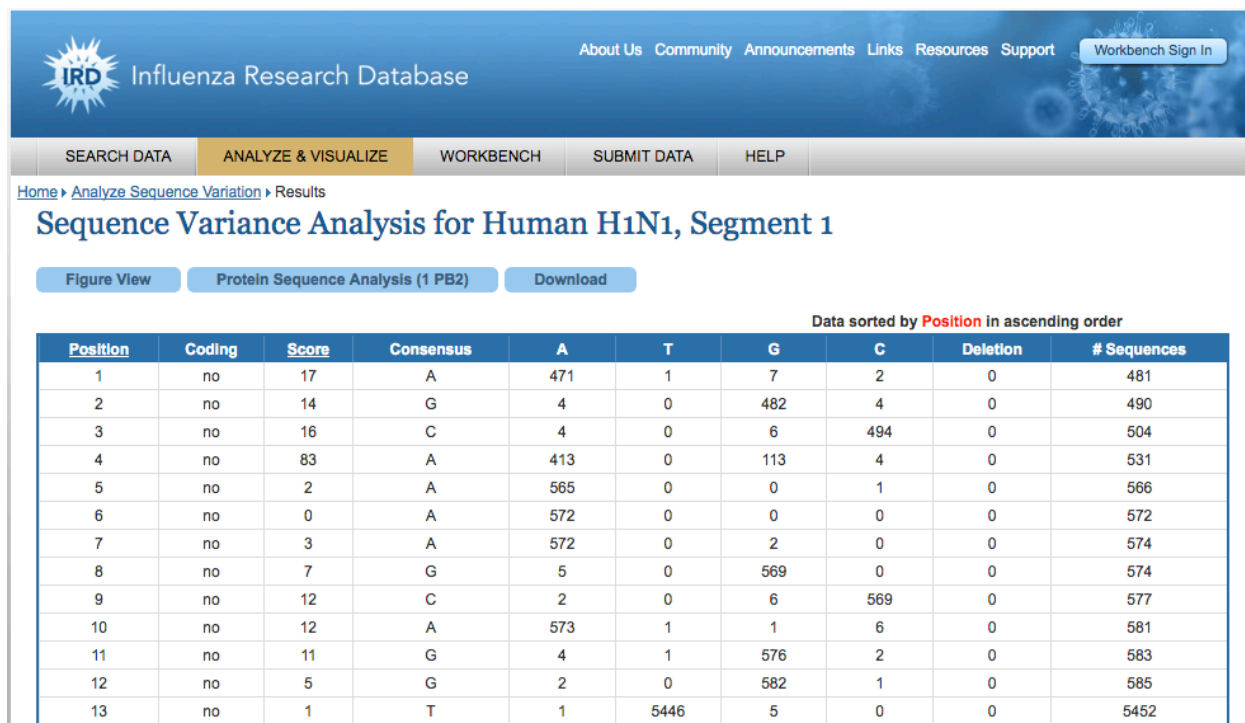
3. Input Data Preparation

Influenza sequences (genomes and proteins) were downloaded from NCBI ftp site at <ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/>. The following files are required: influenza_na.dat, influenza.dat, influenza.faa, influenza.fna, influenza_aa.dat. All sequences were loaded into IRD. All sequences were processed through the IRD curation pipeline to filter out incorrect information and to generate pre-aligned sequences.

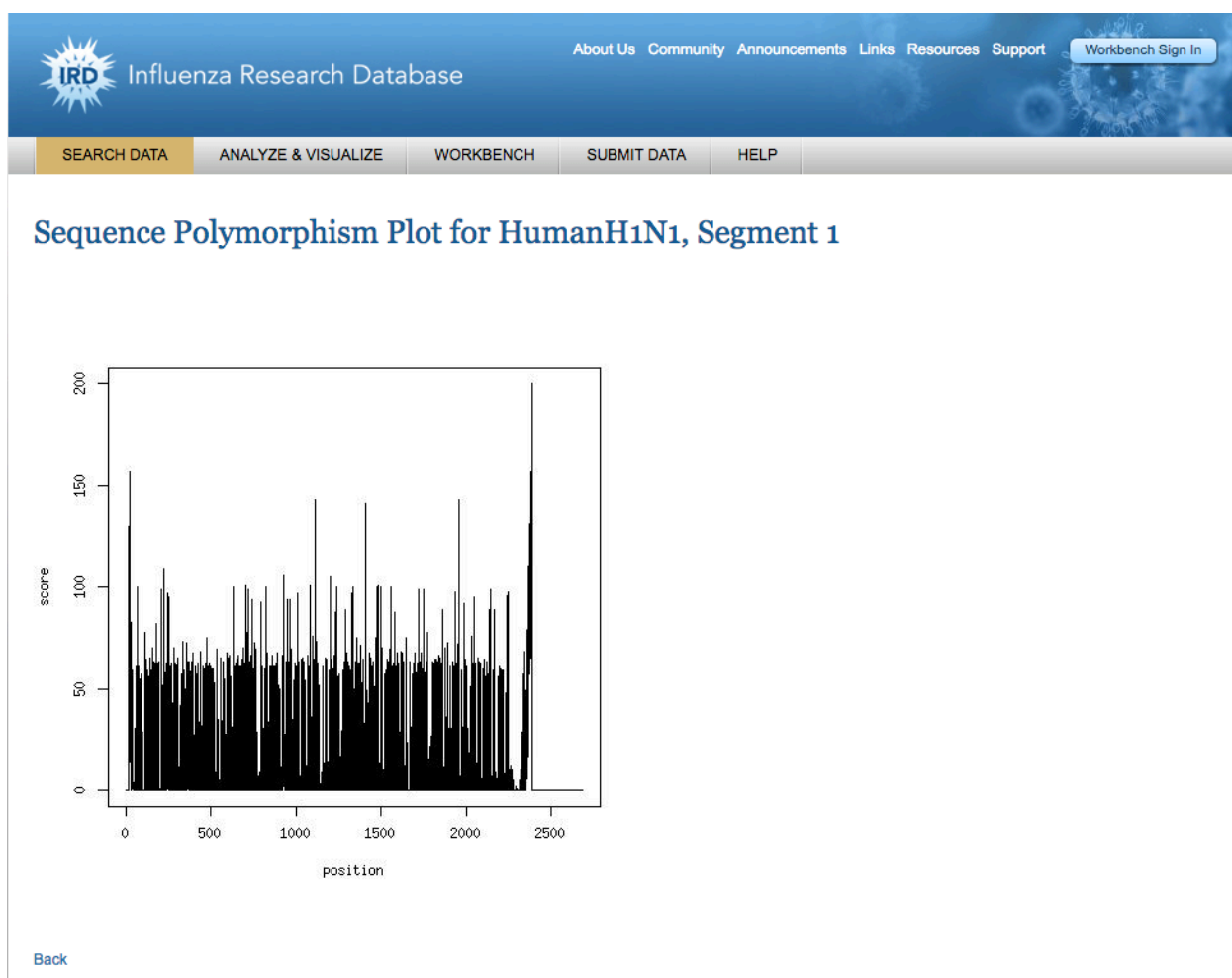
4. Output Data Post-Processing and Display

The final outputs from this pipeline are tab-delimited files for loading into the data warehouse, and tar files that contain the details of alignments and plots for polymorphism scores.


Shown next is an example of the Genomic Sequence Analysis result page for H1N1, Human, Segment 1:



Furthermore, pressing the button “Figure View” yields a plot of variation by position:



Whereas pressing the button “Sequence Analysis” gives a full detail report for every position:


Influenza Research Database

[About Us](#)
[Community](#)
[Announcements](#)
[Links](#)
[Resources](#)
[Support](#)
[Workbench Sign In](#)

[SEARCH DATA](#)
[ANALYZE & VISUALIZE](#)
[WORKBENCH](#)
[SUBMIT DATA](#)
[HELP](#)

[Home](#) > [Analyze Sequence Variation](#) > Results

Protein Sequence Variance Analysis for Human H1N1, Segment 1, PB2

[Genomic Sequence Analysis](#)
[Download](#)

Data sorted by **AA Position** in ascending order

Reference Coordinate

Select from the drop-down list to convert the existing position numbers to a different numbering (coordinate) scheme using a Reference Sequence (e.g. convert from H1 to H3 numbering)

Segment_Number:Protein_Name Strain_Name(Subtype)

| AA Position | Consensus | Score | Alignment Details | # Sequences | Sequence Feature |
|-------------|-----------|-------|--|-------------|-------------------------|
| 1 | Met | 0 | Met=11662 | 11662 | View SF |
| 2 | Glu | 2 | Glu=11666,Gly=13,Trp=1,Val=1,Xaa=1 | 11682 | View SF |
| 3 | Arg | 2 | Arg=11673,Glu=15,Gly=2,Lys=4,Ser=1,Xaa=1 | 11696 | View SF |
| 4 | Ile | 1 | Asp=3,Ile=11705,Met=1,Thr=1,Val=4 | 11714 | View SF |
| 5 | Lys | 1 | Arg=5,Asn=2,Lys=11718,Xaa=3 | 11728 | View SF |
| 6 | Glu | 3 | Arg=10,Asp=3,Gln=1,Glu=11710,Gly=4,Lys=4,Val=1,Xaa=2 | 11735 | View SF |
| 7 | Leu | 1 | Arg=1,Leu=11761,Met=2,Pro=3 | 11767 | View SF |
| 8 | Arg | 0 | Arg=11790,Lys=1,Met=1 | 11792 | View SF |
| 9 | Asp | 60 | Asn=1695,Asp=10101,Gly=3,Ile=1,Ser=1,Tyr=1,Xaa=1 | 11803 | View SF |
| 10 | Leu | 0 | Gln=1,Leu=11802,Met=1 | 11804 | View SF |
| 11 | Met | 0 | Ile=2,Met=11813,Xaa=1 | 11816 | View SF |
| 12 | Ser | 2 | Ala=1,Leu=9,Pro=1,Ser=11801,Thr=1,Trp=3,Xaa=1 | 11817 | View SF |
| 13 | Gln | 1 | Gln=11814,His=4,Leu=2 | 11820 | View SF |
| 14 | Ser | 1 | Phe=6,Pro=5,Ser=11811 | 11822 | View SF |

References

1. Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22), 4673–4680.
2. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res.* 2004 Jun;14(6):1188-90.
3. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004 Mar 19;32(5):1792-7. Print 2004.