1. Purpose

This document describes the mechanism and process by which polymorphism data are generated for influenza sequences stored in the Influenza Research Database (IRD) database.

2. Method Description

The IRD influenza polymorphism data production pipeline is composed of two PERL scripts. All influenza A sequences were downloaded from Genbank. These sequences were processed through the IRD curation pipeline. During this processing, pre-aligned sequences were generated with the ClustalW multiple alignment tool. 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).

$S = -100 * \text{Sum} (Pi * \log Pi)$ where Pi is the frequency of the ith allele

Basically, the score 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 200 (e.g. 4 alleles with 25% frequency each).

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 classified each position as coding or non-coding based on the DNA-protein alignment.

3. Input Data Preparation

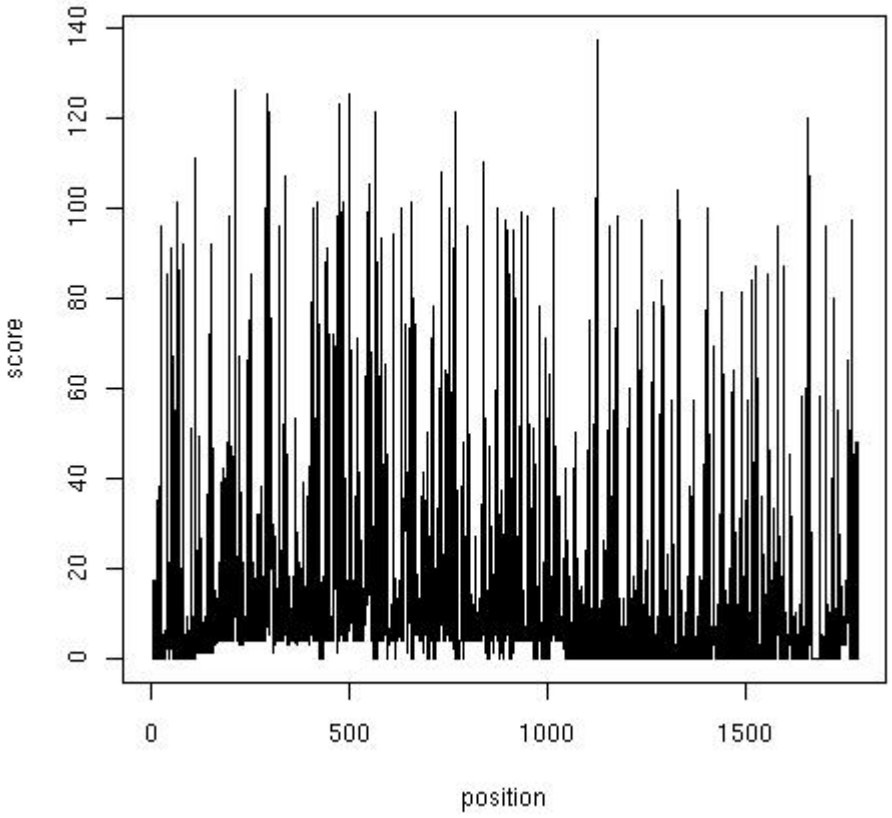
Influenza sequences (genomes and proteins) were downloaded from NCBI ftp site (**ftp.ncbi.nih.gov:/genomes/INFLUENZA/**). The following files are required: influenza_na.dat, influenza.dat, influenza.faa, influenza.fna, influenza_aa.dat. All sequences were loaded into IRD. All sequences were processed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

4. Output Data Post-Processing and Display

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

Below is an example of the Genomic Sequence Analysis result page for H1N1, Human, Segment 4 and the corresponding polymorphism plot:

Position	Coding	Score	Consensus	A	T	G	C	Deletion	# Sequences
1	no	0	A	37	0	0	0	0	37
2	no	0	G	0	0	38	0	0	38
3	no	0	C	0	0	0	38	0	38
4	no	0	A	39	0	0	0	0	39
5	no	17	A	40	0	1	0	0	41
6	no	17	A	40	0	0	1	0	41
7	no	17	A	40	0	0	1	0	41
8	no	17	G	1	0	40	0	0	41
9	no	13	C	0	0	1	54	0	55
10	no	13	A	54	1	0	0	0	55
11	no	0	G	0	0	55	0	0	55
12	no	0	G	0	0	80	0	0	80
13	no	8	G	1	0	95	0	0	96
14	no	35	G	13	0	188	0	0	201
15	no	8	A	201	0	2	0	0	203
16	no	0	A	206	0	0	0	0	206
17	no	0	A	206	0	0	0	0	206
18	no	38	A	192	13	1	0	0	206
19	no	8	T	2	205	0	0	0	207
20	no	14	A	203	4	0	0	0	207
21	no	8	A	216	0	2	0	0	218
22	no	0	A	235	0	0	0	0	235


score
0
20
40
60
80
100
120
140
0
500
1000
1500
position

Below is the Protein Sequence Analysis result page for H1N1, Human, Segment 4, HA protein:

AA Position	Consensus	Alignment Details	# Sequences
1	Met	Met=601	601
2	Lys	Asn=1,Glu=4,Lys=607,Thr=1	613
3	Ala	Ala=451,Thr=1,Val=163	615
4	Lys	Arg=11,Asn=6,Ile=6,Lys=593	616
5	Leu	Arg=1,Leu=622	623
6	Leu	Ile=2,Leu=614,Met=4,Phe=1,Val=3	624
7	Val	Ile=35,Val=593	628
8	Leu	Leu=628	628
9	Leu	Leu=628,Phe=1	629
10	Cys	Cys=620,Ser=1,Tyr=8	629
11	Thr	Ala=266,Pro=1,Ser=1,Thr=361,Val=1	630
12	Phe	Ile=1,Leu=112,Phe=517	630
13	Thr	Ala=23,Ser=92,Thr=515,Val=4	634
14	Ala	Ala=633,Thr=1	634
15	Thr	Ala=16,Leu=1,Thr=618	635
16	Tyr	Asn=9,Asp=156,Lys=1,Ser=2,Tyr=468	636
17	Ala	Ala=636	636
18	Asp	Asn=1,Asp=739,Ile=1	741
19	Thr	Pro=1,Thr=744	745
20	Ile	Ile=740,Leu=6	746
21	Cys	Cys=746	746
22	Ile	Ile=741,Val=5	746

Below is the SNP detail page showing the different SNPs identified between CY009276 and the consensus sequence generated from the H1N1, Human, Segment 4 alignment. This data can be accessed from the Gene Details screen after performing a Sequence search.

SNP ID	Coding	Position in Strain Segment	Position in Consensus	SNP Type	SNP
163410	no	1	15	mismatch	A->G
163411	no	5	19	mismatch	T->A
163412	no	6	20	mismatch	A->T
163413	yes	24	38	mismatch	A->G
163414	yes	29	43	mismatch	A->G
163415	yes	49	63	mismatch	A->G
163416	yes	52	66	mismatch	T->A
163417	yes	55	69	mismatch	A->G
163418	yes	64	78	mismatch	T->G
163419	yes	96	110	mismatch	C->G
163420	yes	124	138	mismatch	G->A
163421	yes	129	143	mismatch	T->C
163422	yes	132	146	mismatch	G->A
163423	yes	165	179	mismatch	A->G
163424	yes	168	182	mismatch	T->C
163425	yes	171	185	mismatch	G->A
163426	yes	177	191	mismatch	T->C
163427	yes	183	197	mismatch	T->C
163428	yes	186	200	mismatch	A->G
163429	yes	190	204	mismatch	C->T
163430	yes	196	210	mismatch	C->A
163431	yes	197	211	mismatch	T->G

5. References

1. **Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004 Mar 19;32(5):1792-7. Print 2004.**
2. **Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004 Jun;14(6):1188-90.**