BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 18 2007, pages 2368–2375
doi:10.1093/bioinformatics/btm354

Phylogenetics

A quantitative genotype algorithm reflecting H5N1 Avian influenza niches

Xiu-Feng Wan[1,*,†], Guorong Chen[1], Feng Luo[2], Michael Emch[3] and Ruben Donis[4]

[1]Department of Microbiology, Miami University, Oxford, OH 45056, [2]School of Computing, Clemson University, Clemson, SC 29634, [3]Department of Geography, University of North Carolina, Chapel Hill, NC 27516 and [4]Influenza Division, Centers for Disease Prevention and Control, Atlanta, GA 30333, USA

Received on January 24, 2007; revised on June 28, 2007; accepted on July 1, 2007

Advance Access publication July 10, 2007

Associate Editor: Keith Crandall

ABSTRACT

Motivation: Computational genotyping analyses are critical for characterizing molecular evolutionary footprints, thus providing important information for designing the strategies of influenza prevention and control. Most of the current methods that are available are based on multiple sequence alignment and phylogenetic tree construction, which are time consuming and limited by the number of taxa. Arbitrarily defining genotypes further complicates the interpretation of genotyping results.

Methods: In this study, we describe a quantitative influenza genotyping algorithm based on the theory of quasispecies. First, the complete composition vector (CCV) was utilized to calculate the pairwise evolutionary distance between genotypes. Next, Hierarchical Bayesian Modeling using the Gibbs Sampling algorithm was applied to identify the segment genotype threshold, which is used to identify influenza segment genotype through a modularity calculation. The viral genotype was defined by combining eight segment genotypes based on the genetic reassortment feature of influenza A viruses.

Results: We applied this method for H5N1 avian influenza viruses and identified 107 niches among 283 viruses with a complete genome set. The diversity of viral genotypes, and their correlation with geographic locations suggests that these viruses form local niches after being introduced to a new ecological environment through poultry trade or bird migration. This novel method allows us to define genotypes in a robust, quantitative as well as hierarchical manner.

Contact: wanhenry@yahoo.com or fvq7@cdc.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Influenza A virus is a negative-stranded RNA virus with eight genomic segments encoding RNA polymerases (PB2, PB1, PA), hemagglutin (HA), nucleoprotein (NP), neuramindase (NA), matrix protein (MP) and non-structural protein (NS). A total of 16 HA (hemagglutinin) and 9 NA (neuraminidase) subtypes have been reported (Fouchier *et al*., 2005; Rohm *et al*., 1996). All of these subtypes come from avian species (Kilbourne, 1997). New influenza viruses and genotypes continuously emerge due to the frequent evolutionary events including genetic reassortment, recombination and mutation. These emerging influenza viruses have caused three pandemics, including H1N1 in 1918, H2N2 in 1957 and H3N2 in 1968 (Webby and Webster, 2003). Influenza genotype analyses reflect influenza viral evolutionary footprints, and thus are critical for preparing a strategy to prevent and control influenza epidemics and pandemics.

Although some bench laboratory methods have been generated for genotyping (Ghindilis *et al*., 2006), these types of approaches can be laborious and time consuming, and thus are not efficient for identifying genetic reassortment events, particularly when a pandemic occurs. A robust and efficient computational genotyping system is crucial; however, genotyping analyses are not trivial tasks. To date, multiple sequence alignments (MSA), [e.g. through Clustal W (Thompson *et al*., 1994) and MUSCLE (Edgar, 2004)], followed by phylogenetic tree construction [e.g. through PAUP* (Swofford, 1998) and Phylip (Felsenstein, 1989)] are the most common approaches used to define genotypes for influenza viruses (Chen *et al*., 2004, 2005, 2006; Guan *et al*., 2000, 2002a, 2002b, 2003, 2004). Since both MSA and tree construction are heuristic approaches, the results will be less reliable as the number of taxa increases. To obtain reliable results, a small group of sequences must be selected as prototypes for the phylogenetic analyses; however, there is no defined reference for this selection. Since we do not know what viruses will emerge in the future, the genotypes assigned are confusing and sometimes conflicting. On the other hand, current influenza genotypes are defined arbitrarily. Even though bootstrap values are generally used to evaluate the confidence of a certain clade or cluster, users can assign two clades as a single genotype or two distinct genotypes. So far, there has not been a standard to define genotypes. Furthermore, the sampling bias through surveillance and sequencing generate another challenge for genotype analysis. We still do not have a good understanding of the influenza virus pool around the world. The World Health Organization (WHO) surveillance lab generally sequences viruses with dissimilar phenotypes. Thus, a single virus with a currently

*To whom correspondence should be addressed.
†Present address: Molecular Virology and Vaccine Branch, Influenza Division, Centers for Disease Prevention and Control, Atlanta, GA 30333, USA.

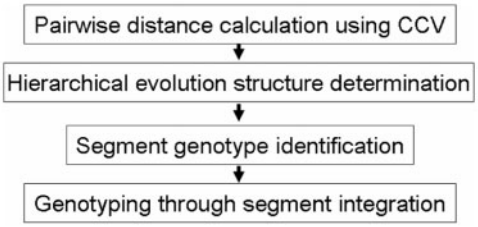

Fig. 1. Simplified workflow of the genotyping method.

defined genotype is more likely to be named with a different genotype in a later analysis when the sequence number increases. Thus, there is a need for a quantitative, robust, and efficient genotyping method.

In 1996, our surveillance isolated two strains of avian influenza viruses (AIV), A/Goose/Guangdong/1/96 (H5N1) and A/Goose/Guangdong/2/96 (H5N1), from sick geese in Shanshui, a small west-central town in Guangdong providence (Guo *et al.*, 1998; Wan, 1998). Since then, the Gs/Gd/96-like H5N1 AIVs have been identified in birds across Asia, Europe and Africa, and at least 317 human cases have been confirmed leading to 191 deaths (http://www.who.int). Recently, familial cases of the virus in Indonesia have presented the possibility of human to human transmission (Kandun *et al.*, 2006). This H5N1 AIV has posed the threat of a future flu pandemic. During the past decade, these viruses have undergone rapid evolution, and multiple genotypes have emerged (Guan *et al.*, 2002). Many influenza genotypes, such as Gs/Gd, A, B, C, D, E, X0, Z, Z+, Y, W, X0–X2 and V (Guan *et al.*, 2004), have been reported. Due to the limitation of current genotyping methods, it is difficult for the scientific community to follow and interpret the results of genotyping analyses.

In this study, we present a novel approach for influenza genotyping based on the theory of quasispecies and under the assumption that influenza A viruses form a niche in response to the environmental pressure. We apply this method in H5N1 influenza genotyping and the results are discussed.

2 MATERIALS AND METHODS

2.1 Algorithm

As shown in Figure 1, we first use the complete composition vector (CCV) to calculate the pairwise evolutionary distance between genotypes. Then, the distribution for each gene segment is assessed with multiple normal distributions using Monte Carlo Bayesian simulation. Through simulation, we are able to find a genotype threshold for each gene segment. With these genotype thresholds, we apply a modularity calculation to identify the influenza segment genotype. At the end, we can combine eight segment genotypes to form a genotype for a virus based on the genetic reassortment feature of influenza A viruses.

2.1.1 Phylogenetic distance calculation To evaluate the evolutionary relationship between influenza A viruses, we need to measure the pairwise distance between gene segments of these viruses. These distances were generally measured by calculating the MSA score (Thompson *et al.*, 1994) or pairwise hamming distance (Plotkin *et al.*, 2002). In this study, a CCV was used to calculate the evolutionary distance between influenza A viruses. This method was described previously (Wan *et al.*, 2007; Wu *et al.*, 2006), and it has been applied in evolutionary study in whole genome sequences (Wu *et al.*, 2006b), influenza reassortment identification (Wan *et al.*, 2007) and HIV genotyping (Wu *et al.*, 2006a; Wu *et al.*, 2007). Briefly, we use S to denote one of the eight gene segments. By using window size k ($k \geq 2$), we can scan S to generate $L-k+1$ strings with the length k; each string can be represented as $\alpha_1\alpha_2,\ldots,\alpha_k$, $\alpha \in$ {A, T/U, G, C} for nucleotide, and α will have 20 amino acids for protein. We denote the number of occurrences in S as $f(\alpha_1\alpha_2,\ldots,\alpha_k)$. Thus, the probability of string $\alpha_1\alpha_2,\ldots,\alpha_k$ is $p(\alpha_1\alpha_2,\ldots,\alpha_k)=f(\alpha_1\alpha_2,\ldots,\alpha_k)/(L-k+1)$. Similarly, we can calculate the probability of string $\alpha_1\alpha_2,\ldots,\alpha_{k-1}$ and $\alpha_2\alpha_3,\ldots,\alpha_k$. Thus, we can calculate the expected probability for string $\alpha_1\alpha_2,\ldots,\alpha_k$ through a Markov model:

$$p^e(\alpha_1\alpha_2,\ldots,\alpha_k)=\begin{cases}\frac{p(\alpha_1\alpha_2,\ldots,\alpha_{k-1})\times p(\alpha_2\alpha_3,\ldots,\alpha_k)}{p(\alpha_2\alpha_3,\ldots,\alpha_{k-1})}, & \text{if} \quad p(\alpha_2\alpha_3,\ldots,\alpha_{k-1})\neq 0\\ 0\end{cases}$$

If $k=1$, we can simplify the above equation as $p^e(\alpha_1)=p(\alpha_1)$. We can calculate the difference $d(\alpha_1\alpha_2,\ldots,\alpha_k)$ between the real probability and expected probability as the evolutionary information carried by string $\alpha_1\alpha_2,\ldots,\alpha_k$:

$$d(\alpha_1\alpha_2,\ldots,\alpha_k)=\begin{cases}\frac{p(\alpha_1\alpha_2,\ldots,\alpha_k)-p^e(\alpha_1\alpha_2,\ldots,\alpha_k)}{p^e(\alpha_1\alpha_2,\ldots,\alpha_k)}, & \text{if} \quad p^e(\alpha_1\alpha_2,\ldots,\alpha_k)\neq 0\\ 0\end{cases}$$

We denote the composition vector containing string $\alpha_1\alpha_2,\ldots,\alpha_k$ as $V^k(S)$, in which k is the length of the string. For a protein sequence which has 20 amino acids, the number of entries in V^k will be 20^k. For the nucleotide sequence, the number of entries in V^k will be 4^k.

To maintain as much information as possible, we include all strings from length m to n, that is $m \leq k \leq n$. Thus, the CCV can be defined as $V(S)=\{V^m(S), V^{m+1}(S),\ldots,V^k(S),\ldots,V^{n-1}(S), V^n(S)\}$. We will determine the optimal values of m and n by checking the amount of information contained. In this study, we set $m=1$ and $n=40$. The distance between two segments is computed using the *Euclidean* distance (Wan *et al.*, 2004):

$$D(S,S')=\sqrt{\sum_{i-1}^{N}(d_i-d_i')^2}$$

where d and d' are the evolutionary information for string $\alpha_1\alpha_2,\ldots,\alpha_k$ in segment S and S' in two AIVs, and $N=20^n$ for amino acids or 4^n for nucleotides, n is the maximum length of the string. For instance, S and S' can be HA segments in two AIVs, respectively.

2.1.2 Quasispecies population landscape identification During Darwinian natural selection, influenza A viruses form quasispecies like other RNA viruses (Domingo *et al.*, 2002). During the host adaptation and disease progression, the fittest virus and a group of variants are selected. Influenza A viruses have been reported to form swarms of selected units in order to adapt to immunological pressure from vaccination (Plotkin *et al.*, 2002). Thus, we hypothesize that influenza A viruses will form a niche in order to adapt to their environment, characterized by such variables as the human or animal (including bird) population and farming systems (including the vaccination program) in a specific geographic area. In order to move from one environment to another with different ecological pressures, the viruses will form a new niche. At the same time, influenza viruses will form 'jumping clusters' in their fitness landscape due to the accumulate of mutations, which is independent of viral population size (Domingo *et al.*, 2002; Plotkin *et al.*, 2002). The distance matrices indicate that each gene segment has a special distribution with different peaks (Fig. 2), each of which contains a mixture of subpopulations. Different peaks show diversity at different levels for the number of mutation accumulations.

In order to dissect the diversity within the population landscape, each gene segment matrix was analyzed using Bayesian inference Using

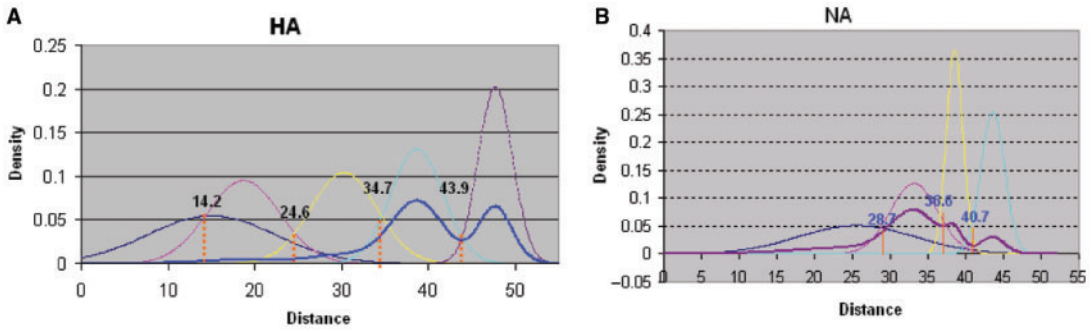

Fig. 2. The mixtures of subpopulations in HA and NA gene segment. The distributions were estimated using hierarchical Bayesian model by Bayesian inference using Gibbs Sampling method through BUGS version 1.3, software developed by MRC Biostatistics Unit. (**A**) HA gene segment; (**B**) NA segment.

Gibbs Sampling (BUGS) software developed by MRC Biostatistics Unit. The Hierarchical Bayesian (HB) model was employed to fit our data:

$$y_i \sim \text{Normal}(\mu_{T(i)}, \tau^{-1}_{T(i)}), \text{ where}$$

y_i is the ith data point; $T(i)$ is from a categorical distribution with category probabilities $\mathbf{P}=(p_1,p_2,\ldots,p_J)^T$ and $\sum_j p_j=1$ $(j=1,2,\ldots,J)$; Vector $\mathbf{P}$ has a Dirichlet distribution with parameter $\boldsymbol{\alpha}=(1,1,\ldots,1)^T_{J\times 1}$; Probability p_j stands for the proportion of data points from the jth normal distribution. To avoid all data points going into one group, the following restriction was added: $\mu_1 < \mu_2 <,\ldots,< \mu_J$, where J is the total number of mixing components. J is predetermined based on the histogram of our data. All μ's and τ's are given independent non-informative prior distributions. After the above model was developed, each normal distribution with this mixing distribution is known. *The* jth *niche point* is defined as the corresponding data point where its jth density is equal to its $(j+1)$th density. These niche points are called genotype thresholds and are used to dissect influenza niches. To get robust parameter estimates the first 3000 runs are thrown away. All parameters are estimated based on the remaining 2000 runs per gene segment.

2.1.3 Population structure representation and genotype identification Influenza A viruses can be viewed as the space of an interwoven graphic network, with each node in the network representing a virus (Huynen *et al.*, 1996). Influenza niches will form a distinct clique in this network, and the genotype threshold will help dissect the cliques in this graph. In some cases, however, niches may be linked as a loose clique. Thus, we need to further apply the module identification algorithm to isolate these modules into a clique.

A local optimization algorithm is used to identify modules within an un-weighted and un-directed graph (Luo *et al.*, 2006). Given a vertex v, a subgraph S with v is created. Then, vertices adjacent to v are placed in its neighbor set N. This algorithm is recursive with *addition* and *deletion* steps. In the *addition* step, the vertices from N are iteratively added to the subgraph S. In the *deletion* step, a vertex is removed from S if it has more edges in another subgraph S'. In the end, all vertices are assigned into different subgraphs. A subgraph S is defined as a module if modularity M is more than a threshold of 1. M is calculated as

$$M=\frac{\text{ind}(S)}{\text{outd}(S)}$$

where the ind(S) is the total number of edges within the subgraph S, and outd(S) is the total number of edges outside subgraph S. A node i will be defined as an outlier of module α if

$$d^{\min}_{i,\alpha} = \bar{d}_\alpha + 2\delta_\alpha$$

$d^{\min}_{i,\alpha}$ is the minimum distance of node i to any nodes in the module α, $\bar{d}_\alpha$ is the average evolutionary distance within the module α, and δ_α is the standard deviation of the evolutionary distances within the module α.

Each of these resulting modules will be assigned as a segment genotype. Based on the genetic feature of reassortment, we can assign an influenza genotype by combining eight segment genotypes. Biolayout (Enright and Ouzounis, 2001) is applied to visualize the genotypes we will identify.

2.2 Genotype tree construction

Since each virus has eight segment genotypes, we measure the distance between each virus genotype using the hamming metric

$$D(V_\alpha, V_\beta) = \sum_{i=1}^{N} d_i, \quad N=8$$

where D is the distance between influenza virus α and influenza virus β, N is the total number of gene segments and d_i is the distance between a segment genotype between these two viruses. The value of d_i is 0 if the two segment genotypes are the same, 0.5 if overlapped, and 1, otherwise. The genotype tree was constructed using Neighbor-Joining method in Phylip 3.65 package (Felsenstein, 1989).

2.3 Datasets

The influenza datasets were downloaded from Influenza Virus Resource database at GenBank (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html), which were updated in November 2006. In this study, we analyzed the influenza A viral segments with full length from a pool of 37677 influenza sequences, including 2641 PB2, 2630 PB1, 2650 PA, 963 H5 of HA, 3003 NP, 1120 N1 of NA, 3474 MP and 3098 NS full genomic segments. The final genotyping analyses were focused on 283 H5N1 AIVs with eight full-length genomic segments.

3 RESULTS

3.1 HA population landscape of H5 AIVs

Like other RNA viruses, influenza A viruses form quasispecies through accumulative mutations, and possibly recombinations, during their adaptation when there is environmental pressure. The distance analysis of all H5 gene segments showed five peaks in their distance distribution. By Bayesian inference using

Gibbs Sampling method, we were able to identify the mixed hierarchical evolution structures for H5 HA gene segment. HA gene has five mixed normal distributions with four genotype thresholds, 43.9, 34.7, 24.6 and 14.2, respectively. We denote these five peaks as peak 1–5 (Fig. 2). Since these distances are pairwise distances between influenza A viral segments, the higher the genotype threshold value the less separation between viruses. These peaks can be ranked in a hierarchical structure, with peak 1 a 'looser' adaptation, and the fifth peak a 'tighter' one towards to an adaptation to a more specific environment. Each of these peaks is composed of a number of subpopulations, and each subpopulation will be an assemblage of influenza A viruses with closer evolutionary distance. At each level, a subpopulation may be considered a genotype or niche. Each genotype threshold reflects a 'jumping' adaptation during the Darwin neutral selection (Plotkin *et al.*, 2002). We rank genotype thresholds from high to low as Level 1 (L1) to Level 4 (L4), and use them as bases for hierarchical segment genotyping method described later.

To determine whether the matrix distribution is a general phenomenon for H5 HA population, we simulated HA population landscape by randomly selecting 200 sequences from the current H5 HA gene pool. The results demonstrated that the randomly sampled segments and the overall H5 HA have a similar distribution (see Supplementary Fig. 1). It is also shown that the viruses from a single genotype with large enough data size (e.g. viruses from migration birds) showed a similar matrix distribution (data not shown). Thus, the peak boundary will serve as a baseline for a robust and quantitative genotyping method.

3.2 Hierarchical segment genotyping method

Using traditional computational genotyping methods, an influenza genotype is defined as a clade/cluster of viruses among phylogenetic tree. A common feature for a genotype is that the viruses within a genotype are phylogenetically closer to each other than the viruses outside this genotype. By considering this definition, we are able to develop a genotype definition using module identification within a network. By assuming the inner degree within a module is larger than its other degree, we can even further define the detailed groups among H5 influenza A viruses. This will form a basis for us to define a niche or a genotype. By dissecting the distance matrix of influenza viruses at a genotype threshold, we can separate these viruses into different subgroups. Through graphical viewing, we can consider this dissection process for separating the influenza viruses into different modules.

Using the hierarchical structural relationship, we were able to genotype H5 influenza at four different levels (Fig. 3). A genotype in a higher level may include multiple genotypes in a lower level. By using the threshold of 43.9 (L1), seven HA segment genotypes were identified and only four of them have 10 or more viruses. At the level of 34.7 (L2), 28 HA genotypes were identified, 6 of which have at least 10 viruses. Using the threshold of 24.6 (L3), we identified 65 HA genotypes and 19 of them have at least 10 viruses. At the level of 14.2 (L4), 116 HA genotypes were identified and 16 of these genotypes have at least 10 viruses.

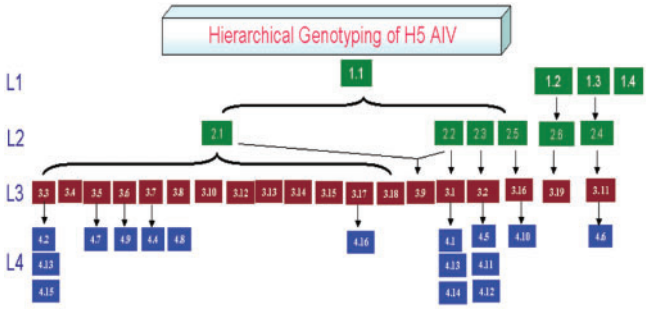

Fig. 3. Hierarchical genotyping results for H5 HA gene of influenza A viruses. Only the genotypes with at least 10 viruses are shown. Different layer shows the genotypes using different genotype threshold, which is estimated through Bayesian influence based on the evolutional distance between viruses.

Similar to HA, we identified four to five peaks, and thus three to four genotype thresholds for the other seven gene segments: PB2 (59; 52.4; 38.9), PB1(54.4; 46.4; 34.1; 21.9), PA(47.4; 34.9; 25.5), NP(43.8; 38.3; 25.7), NA(40.7; 36.6; 28.7), MP(29.1;21.4;13.9) and NS(34.4; 27.5; 20.2) (Supplementary Fig. 2). Among these eight gene segments, only HA and PB1 have five peaks, and this suggests that both genes of H5N1 AIVs are more dynamic than the other six segments and form subpopulations in order to adapt to local environments. The HA gene is a surface glycoprotein which is a major target for host immune responses. In order to evade antibody neutralization, the HA gene must mutate quickly. The PB1 gene is responsible for transcription initiation and cRNA synthesis (Steinhauer and Skehel, 2002). Recently, PB1-F2 was reported to be important in viral pathogenesis. Both PB1 and HA were transferred in their entirety from avian species to the 1957 and 1968 pandemic strains (Belshe, 2005). PB1-F2 genes were shown to be under positive selection in recent Indonesian epidemics (Smith *et al.*, 2006b). Though it is not clear why PB1 has a different distance distribution than the other two polymerases, we speculate that this might be associated with the host transmission from avian species to humans.

In this research project, we used L3 to perform the genotype analysis since this genotype threshold was able to remove more than 95% of the vertices in the network. We compared our genotype analyses with traditional genotype analyses using phylogenetic tree. By comparing these groups and the clusters in the phylogenetic tree, we found that these subgroups correspond well to clades in phylogenetic trees. Figure 4 shows the module representation of 27 genotypes with at least five nodes. Some genotypes are still interconnected; for instance, a big cluster C-I showed the linking traces from 3.15 (Gs/Gd-like) and China Mixture genotypes (3.13, 3.4, 3.8) to Hong Kong 02/03 (3.9) and Vietnam/Thailand (3.1). Two nodes with maximum linkages from each genotype with at least five nodes were selected to construct a phylogenetic tree (Supplementary Fig. 3), and it showed an evolutionary relationship between these HA segment genotypes. Our genotyping method first quantitatively dissects genotypes. A previously reported genotype may be classified as multiple genotypes in our analysis, and vice versa;

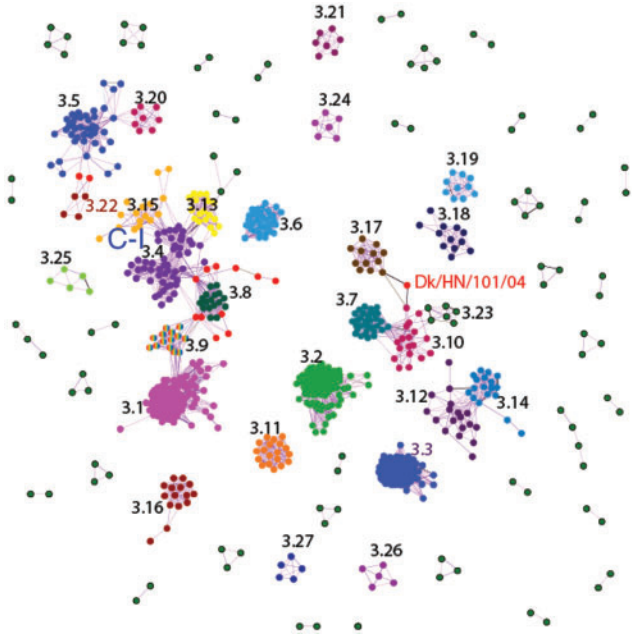

Fig. 4. Segment genotyping results of H5 AIV HA gene segment. The module representation of genotypes. Each genotype was marked as an individual module through 3.1–3.27. Only the modules with at least five nodes are marked. The edge between two nodes reflects their evolutionary distance is smaller than the segment genotype threshold (the length of the edge is not correlated with the evolutionary distance calculated from CCV).

for instance, Gs/Gd-like (3.15) and China Mixture (3.13, 3.4, 3.8 and 3.9) were grouped as a single genotype Gs/Gd-like for HA gene (Guan *et al*., 2004).

During evolution, some influenza A viruses serve as intermediate strains between niches and may be a beginning point for a new niche initiation. Being reflected in segment genotype identification, some genes can belong to more than one genotype. These genes will be assigned with multiple segment genotype numbers; for instance, Dk/HN/101/04 may be assigned with genotypes 3.10 and 3.17 since it is linked with both of these two genotypes (Fig. 4 and Supplementary Fig. 4).

Similar to HA segment, we use the L3 threshold to separate subgroups for these seven gene segments and then apply module identification algorithm. As a result, we obtained 27 segment genotypes for PB2, 17 for PB1, 33 for PA, 24 for NP, 9 for NA, 44 for MP and 31 for NS. The genotyping results for all eight gene segments are shown in Supplementary Table 1.

3.3 Reassortment identification

Using similar segment genotype thresholds, we were able to identify reassortments for six internal gene segments (Supplementary Table 2). The genotypes with potential reassortment events are shown in Supplementary Table 3. Our analyses showed that H5N1 AIVs actively reassort with H2N2, H3N8, H6N1, H6N8, H9N2, H9N6, H6N1, H7N1, H11N2 and H11N3 AIV. The majority of these events are between H9N2 and H5N1 AIVs. Since 1994, H9N2 AIVs have been a common epidemic disease in the Chinese poultry industry (Wan, 1998). The prevalence of H9N2 strains provides a continuous donor resource for H5N1 AIV. The evidence shows that the H9N2 viruses are also acceptors of H5N1 AIV. These two directional interactions will facilitate selection of versatile influenza strains. More important, H9N2 has been reported to infect humans, and thus may be a potential pandemic threat (Guan *et al*., 1999, 2000). Our results also show that some H5N1 AIVs have picked up NP gene segments from the H6N2 ad H2N2 AIVs isolated in Hong Kong about 30 years ago. Similarly, NP of Ck/Hebei/718/01 has had a potential reassortment with Africanstarling/England-Q/983/1979(H7N1), and genotype 3.1 (including the Vietnam/Thailand, Indonesia, Japan/Korea and China Mixture) has had a potential reassortment with H9N6 virus isolated from Hong Kong in 1977. It is unknown how these NP segments sustain low evolutionary rates over such a long time. Since our knowledge about the influenza gene segment pool is limited, our analyses are still sporadic. We believe many reassortments have not been identified due to the limitations of influenza sequence data. Nevertheless, our methods provide an efficient approach to identify reassortment events.

3.4 Influenza genotype identification

Based on the feature of the segmented RNA genome, eight segment genotypes are integrated to interpret the genotypes for each virus. In this study, our analyses only focus on the H5N1 AIV with complete genomes. Each virus genotype is assigned a combining number of eight segment genotypes PB2-PB1-PA-HA-NP-NA-MP-NS. As a result, we were able to assign 107 genotypes for 283 H5N1 AIVs with full genome sequences (Supplementary Table 4). For instance, 3.1-3.1-3.1-3.1-3.1-3.2-3.7-3.1 means PB2 segment genotype is 3.1, PB1 3.1, and so on. If a virus was not grouped as a segment genotype with any other virus, its genotype will be assigned as NM.

Compared to previous reported genotypes, our results are more specific and they are able to differentiate genotypes in a quantitative manner. In addition, the presentation of our genotyping method clearly displays the relationship between their segment genotypes. Some previously defined genotypes may be interpreted differently in our analyses. For instance, previously reported Z and Z+ genotype (Li *et al*., 2004) are separated out as more than 20 genotypes in our results, which reflect adaptations of these viruses as niches in different environments. Even for the influenza viruses isolated from Vietnam/Thailand during 2004–2006 outbreaks, there are at least four genotypes produced by our methodology.

The genotype tree clearly shows the relationship between genotypes. For instance, it has been reported that the 97-like H5N1 influenza was detected from eggs of Vietnamese waterfowl (Li *et al*., 2006), which was shown in the Vietnam I genotype cluster (Fig. 5). However, these viruses evolved with the involvement of a local NP gene segment pool, probably the prevailing viruses, as shown in Vietnam IV genotype cluster. A/Duck/Vietnam/1/2005 (3.14-3.4-3.5/3.11-3.12-3.1-3.4-3.42-3.8) might have obtained its NP gene from this genotype cluster IV, while A/Goose/Vietnam/3/2005 (genotype 3.14-3.4-NM-3.12-3.3/3.4-3.4-NM-3.8) and A/Duck/Vietnam/8/2005 (3.14-3.4-NM-NM-3.3/3.4-3.4-3.42-3.8) maintain their NP

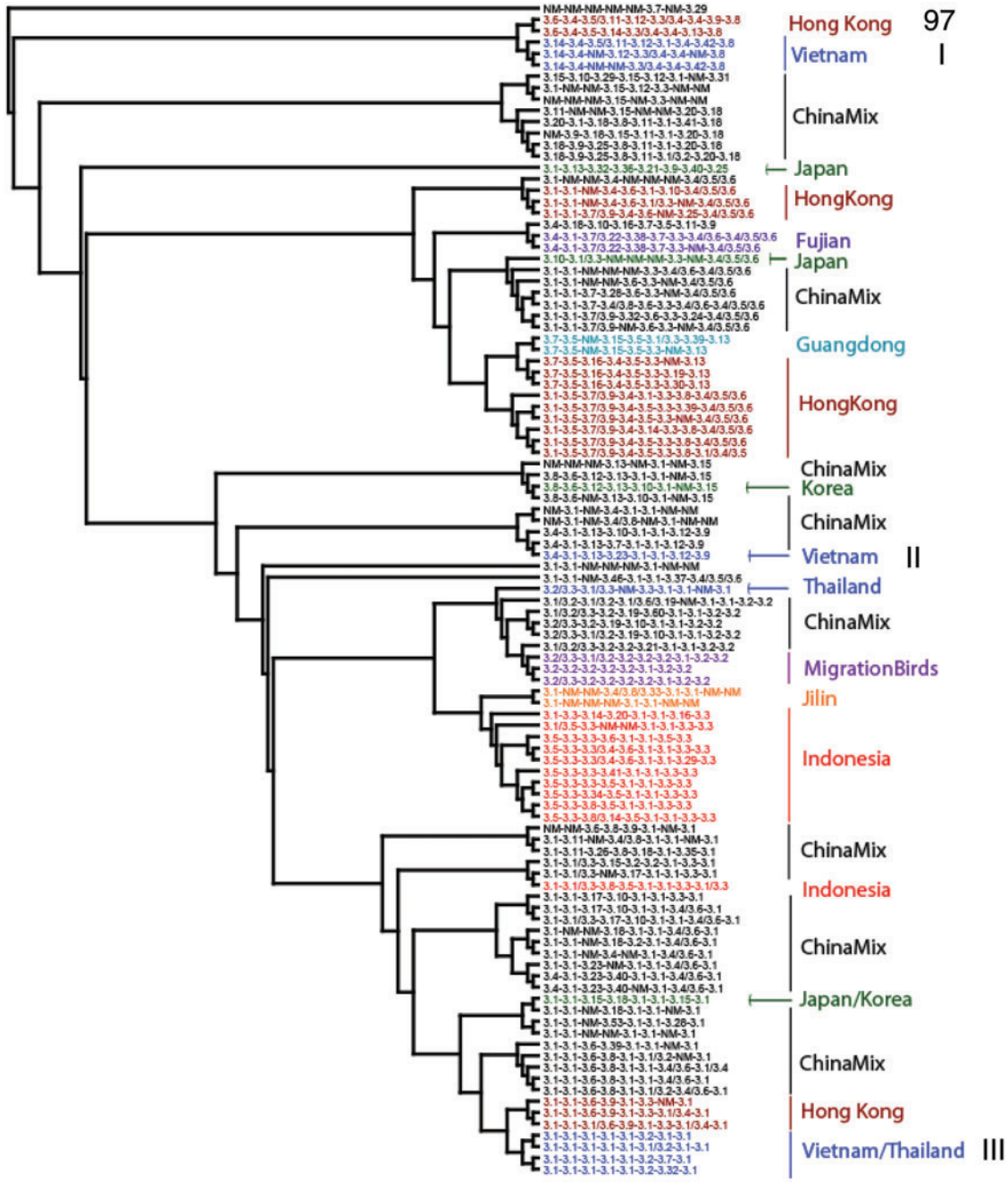

Fig. 5. The genotype tree of H5N1 influenza viruses was constructed through Neighbor-Joining method in Phylip 3.65 package (Felsenstein, 1989) based on the hamming distance between each virus genotype.

gene segment from HK 97-like virus; the phylogenetic tree is shown in Supplementary Figure 5.

3.5 Geographic niches reflected by our genotyping method

The genotype tree in Figure 5 shows the geographic distribution of these genotypes and the evolution pathways of H5N1 AIVs. The continuous isolation of AIVs indicate that China has been the epicenter of H5N1 AIVs. In China, many genotypes were identified in the past 10 years, and many of these genotypes have become local strains interwoven with the local virus pool (Guan *et al.*, 2004; Li *et al.*, 2006). These H5N1 AIVs have been actively reassorted with other influenza viruses, especially H9N2 AIVs (Guan *et al.*, 1999). Through these adaptations, some viruses may form new niches or be replaced by other viruses that have been introduced. Several genotypes are identified with Chinese provinces including Guangdong, Fujian and Jilin as shown in Figure 5, while others are simply

marked as China Mixture. Our genotyping results show that the Vietnam I and III genotypes were related to Hong Kong 97 and 2002 genotypes, respectively. The Vietnam II genotype is closely associated with the Hunan/Guangxi genotypes in China. We identified at least 10 Indonesian genotypes, which are associated with different geographic locations, even within that country. These genotypes may be first introduced by migrating birds and then complicated by domestic poultry movements (Smith *et al*., 2006b).

During transmission through migrating birds, the viruses formed three different genotypes. Each of these genotypes reflects the geographic area among the flyways of migration birds which transmit and spread them (Supplementary Table 4). The first genotype was identified in China, Mongolia, Astrakhan (Russia), Novosibirsk (Russia), Kurgan (Russia), Croatia, Iran, Italy, Germany and Nigeria. The second genotype was identified in the Slovenia, Egypt, Sudan, Nigeria and Ivory Coast, and the third in Afghanistan.

Although the factors associated with this selection are not clear, the different environmental factors in different geographic areas may be linked to the formation of these emerging influenza genotypes. One common factor may be the differences between bird species range, population size, farming system and the structure of the poultry industry. On the other hand, at some locations, multiple genotypes may co-exist, and the local influenza virus pool would facilitate the emergence of new genotypes.

4 DICUSSION

There are at least three advantages of the novel genotyping method presented in this article. First, it avoids MSA and phylogenetic tree construction utilized by most of the currently available methods. The heuristic results from MSA and phylogenetic tree construction are often confusing. Previous studies show most of the available MSA methods only have accuracy of <70%, and even as low as 40% (Lassmann and Sonnhammer, 2002). Even MSA would not be a major problem for influenza viruses but it has been always an issue to refine the alignments in the gap inserted regions, such as variation in HA cleavage site and NA stalk region. Furthermore, phylogenetic trees are not able to generate a reliable result if the influenza sequence number is large. Thus, researchers have to select a subset of sequences for analysis, which will potentially generate a biased result. Second, our method will define genotypes based on the genotype thresholds, which was derived from the influenza viral population. This approach allows us to quantitatively dissect viruses into different subgroups. Third, through module identification we are able to identify the viruses closely correlated with each other; in turn, these viruses are called a module. A module reflects a cluster/genotype of influenza A viruses with close evolutionary distances. These viruses are shown to be biological niches formed through selection from ecological factors. Thus, our methods will be more robust than traditional genotyping methods; they can handle a large number of influenza viruses. The development of quantitative methods makes these processes easier for public heath workers, even those without a background in molecular biology.

The distance distribution of influenza viruses provides us standards for genotyping. Different segment genotype thresholds reflect different levels of adaptation. The smaller a genotype threshold, the more specific the adaptation is to local ecological factors and these factors may include the poultry industry, vaccination programs and any other geographic factors. In a simple way, the hierarchical structure may reflect the overlapping structures similar to the structures of geographic areas. Frequent human activities, however, such as trading and marketing, and migration of birds have made this very complicated. This algorithm allows us to define the genotype in a hierarchical method. The hierarchical structure may be similar to the topology structure of gene ontology (http://www.geneontology.org/). Creating a systematic genotype ontology study is our future goal.

To evaluate our genotyping results, we compared the results from our method with the clades identified in phylogenetic tree using maximum parsimony (Supplementary Fig. 6). Our results showed that most of the genotypes we identified correlate well with the clades shown in the phylogenetic tree. We also concatenated eight segments as a single genome and then performed maximum parsimony tree construction using these concatenated whole flu genomes (Supplementary Fig. 7). It was also shown that most of the genotypes from our methods correlate well with the clades shown in whole genome phylogenetic tree. Those viruses with reassortments were shown to be distinct clades, which were also assigned as different genotypes in our methods. Principle coordinate analysis (PCoA) has been applied in traditional genotyping (Malysheva-Otto *et al*., 2006). We applied PCoA using the distance matrices from CCV, and the results demonstrated that PCoA only distinguished major genotypes our method identified (Supplementary Fig. 8). Similar to phylogenetic tree construction, it is difficult to define the clusters' boundary when using PCoA. Thus, we believe our methods are more robust than both phylogenetic tree construction and PCoA in influenza genotyping.

The drawback of our method is that current distance measurement is somewhat dependent on the length of sequences. Thus, our methods require complete or nearly complete genomic sequences. Presently, we believe that the advances in genomic sequences have compensated for these drawbacks. On the other hand, the influenza sequencing project launched by National Institute of Allergy and Infectious Diseases (NIAID) in November 2004 have dramatically decreased this limitation.

The development of this method is based on the theory of quasispecies. Like other RNA viruses, influenza A viruses form quasispecies during evolution. During natural selection, the fittest sequence and an assemblage of different variants are chosen (Domingo *et al*., 2002). The fittest sequence may represent only a very small fraction of the population. The quantitative dissection may facilitate the identification of this fittest genotype from other subpopulations. In contrast, through traditional genotyping methods, it will be easy to confuse this 'orphan' genotype with other genotypes in the first stage of an outbreak, and this 'orphan' genotype may only be identified after becoming domain strains (Smith *et al*., 2006a).

Influenza A viruses can be viewed as a space of an interwoven network (Huynen *et al*., 1996), and these viruses evolve as swarms of selective units (Plotkin *et al*., 2002). Thus, integration of segment genotypes reflects the complicated relationship between genotypes within this network. The combination process is similar to the HA and NA serotype combination but much more specific. By applying our method to the H5N1 genotyping, we identified more than 100 niches in 283 H5N1 AIVs with complete genomes. Our methods show that a virus forms new niches after being introduced into a new geographic location through migrating bird or poultry movement (Kilpatrick *et al*., 2006). The driving forces for these emerging genotypes may be caused by the local ecological factors and human activities, such as animal populations, influenza virus pool, and vaccination strategies. As a result, our genotyping method may provide a useful method to study the evolutionary pathway of influenza viruses.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Yi Guan, Guohui Lin, Helen Piontkivska, Gavin Smith and Linda Waston for their critical discussions. The authors are indebted to Changzheng Rao for assistance of Message Passing Interface implementation of parallel computing. The authors thank computing supports from Jaime Combariza and the research computing support group at Miami University, and from Terry Lewis and the Ohio Supercomputer Center. This work was supported by Miami University CFR grant.

Conflict of Interest: none declared.

REFERENCES

Belshe,R.B. (2005) The origins of pandemic influenza – lessons from the 1918 virus. *N. Eng. J. Med.*, **353**, 2209–2211.

Chen,H. *et al.* (2004) The evolution of H5N1 influenza viruses in ducks in southern China. *Proc. Natl Acad. Sci. USA*, **101**, 10452–10457.

Chen,H. *et al.* (2005) Avian flu: H5N1 virus outbreak in migratory waterfowl. *Nature*, **436**, 191–192.

Chen,H. *et al.* (2006) Establishment of multiple sublineages of H5N1 influenza virus in Asia: implications for pandemic control. *Proc. Natl Acad. Sci. USA*, **103**, 2845–2850.

Domingo,E. *et al.* (2002) *Quasispecies and RNA Virus Evolution: Principles and Consequences*. Landes, Austin, TX.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Enright,A.J. and Ouzounis,C.A. (2001) BioLayout – an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, **17**, 853–854.

Felsenstein,J. (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.

Fouchier,R.A. *et al.* (2005) Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls. *J. Virol.*, **79**, 2814–2822.

Ghindilis,A.L. *et al.* (2007) CombiMatrix oligonucleotide arrays: genotyping and gene expression assays employing electrochemical detection. *Biosens. Bioelectron*, **22**, 1853–1860.

Guan,Y. *et al.* (1999) Molecular characterization of H9N2 influenza viruses: were they the donors of the ''internal'' genes of H5N1 viruses in Hong Kong? *Proc. Natl Acad. Sci. USA*, **96**, 9363–9367.

Guan,Y. *et al.* (2000) H9N2 influenza viruses possessing H5N1-like internal genomes continue to circulate in poultry in southeastern China. *J. Virol.*, **74**, 9372–9380.

Guan,Y. *et al.* (2002a) Emergence of multiple genotypes of H5N1 avian influenza viruses in Hong Kong SAR. *Proc. Natl Acad. Sci. USA*, **99**, 8950–8955.

Guan,Y. *et al.* (2002b) H5N1 influenza viruses isolated from geese in Southeastern China: evidence for genetic reassortment and interspecies transmission to ducks. *Virology*, **292**, 16–23.

Guan,Y. *et al.* (2003) Reassortants of H5N1 influenza viruses recently isolated from aquatic poultry in Hong Kong SAR. *Avian Dis.*, **47**, 911–913.

Guan,Y. *et al.* (2004) H5N1 influenza: a protean pandemic threat. *Proc. Natl Acad. Sci. USA*, **101**, 8156–8161.

Guo,Y. *et al.* (1998) Genetic characterization of an avian influenza A (H5N1) virus isolated from a sick goose in China. *Zhonghua Shi Yan He Lin Chuang Bing Du Xue Za Zhi*, **12**, 322–325.

Huynen,M.A. *et al.* (1996) Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl Acad. Sci. USA*, **93**, 397–401.

Kandun,I.N. *et al.* (2006) Three Indonesian clusters of H5N1 virus infection in 2005. *N. Engl. J. Med.*, **355**, 2186–2194.

Kilbourne,E.D. (1997) Perspectives on pandemics: a research agenda. *J. Infect. Dis.*, **176** (Suppl. 1), S29–S31.

Kilpatrick,A.M. *et al.* (2006) From the cover: predicting the global spread of H5N1 avian influenza. *Proc. Natl Acad. Sci. USA*, **103**, 19368–19373.

Lassmann,T. and Sonnhammer,E.L. (2002) Quality assessment of multiple alignment programs. *FEBS Lett.*, **529**, 126–130.

Li,K.S. *et al.* (2004) Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature*, **430**, 209–213.

Li,Y. *et al.* (2006) Detection of Hong Kong 97-like H5N1 influenza viruses from eggs of Vietnamese waterfowl. *Arch. Virol.*, **151**, 1615–1624.

Luo,F. *et al.* (2006) Exploring local community structures in large networks. In *Proceedings of this 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. Hong Kong, China, pp. 233–239.

Malysheva-Otto,L.V. *et al.* (2006) Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (Hordeum vulgare L.). *BMC Genet.*, **7**, 6.

Plotkin,J.B. *et al.* (2002) Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proc. Natl Acad. Sci. USA*, **99**, 6263–6268.

Rohm,C. *et al.* (1996) Characterization of a novel influenza hemagglutinin, H15: criteria for determination of influenza A subtypes. *Virology*, **217**, 508–516.

Smith,G.J. *et al.* (2006a) Emergence and predominance of an H5N1 influenza variant in China. *Proc. Natl Acad. Sci. USA*, **103**, 16936–16941.

Smith,G.J. *et al.* (2006b) Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam. *Virology*, **350**, 258–268.

Steinhauer,D.A. and Skehel,J.J. (2002) Genetics of influenza viruses. *Annu. Rev. Genet.*, **36**, 305–332.

Swofford,D.L. (1998) *PAUP*: Phylogenic Analysis Using Parsimony*. Sinauer, Sunderland, Massachusetts.

Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Wan,X.-F. (1998) Isolation and Characterization of Avian Influenza Viruses in China. *College of Veterinary Medicine*. South China Agricultural University, Guangzhou.

Wan,X.F. *et al.* (2004) Revealing gene transcription and translation initiation patterns in archaea, using an interactive clustering model. *Extremophiles*, **8**, 291–299.

Wan,X.-F. *et al.* (2007) Computational identification of reassortments in avian influenza viruses. *Avian Dis.*, **51**, 434–439.

Webby,R.J. and Webster,R.G. (2003) Are we ready for pandemic influenza? *Science*, **302**, 1519–1522.

Wu,X. *et al.* (2006a) Whole genome composition distance for HIV-1 genotyping. In Xu,Y. (ed.) *Proceedings of the IEEE Computational Systems Bioinformatics*. Stanford, California, pp. 179–190.

Wu,X. *et al.* (2006b) Phylogenetic analysis using complete signature information of whole genomes and clustered Neighbour-Joining method. *Int. J. Bioinformat. Res. Appl.*, **2**, 219–248.

Wu,X. *et al.* (2007) Nucleotide composition string selection in HIV-1 subtyping using whole genomes. *Bioinformatics*., http://bioinformatics.oxfordjournals.org/cgi/reprint/btm248v1