



ELSEVIER

Contents lists available at ScienceDirect

## Journal of Theoretical Biology

journal homepage: [www.elsevier.com/locate/jtbi](http://www.elsevier.com/locate/jtbi)

# Viral genome phylogeny based on Lempel–Ziv complexity and Hausdorff distance



Chenglong Yu<sup>a</sup>, Rong Lucy He<sup>b</sup>, Stephen S.-T. Yau<sup>c,\*</sup>

<sup>a</sup> Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, IL 60607, USA

<sup>b</sup> Department of Biological Sciences, Chicago State University, IL 60628, USA

<sup>c</sup> Department of Mathematical Sciences, Tsinghua University, Haidian District, Beijing 100084, PR China

## HIGHLIGHTS

- To apply Lempel–Ziv complexity to define the distance between two sequences.
- Use Hausdorff distance (HD) to analyze multi-segmented viral genomes.
- Use a modified Hausdorff distance (MHD) to analyze multi-segmented viral genomes.
- Take the multi-segmented genome as an entirety to make the comparative analysis.

## ARTICLE INFO

### Article history:

Received 5 October 2013

Received in revised form

18 December 2013

Accepted 18 January 2014

Available online 29 January 2014

### Keywords:

Single-segmented

Multi-segmented

Global comparison

Virus classification

## ABSTRACT

In this paper, we develop a novel method to study the viral genome phylogeny. We apply Lempel–Ziv complexity to define the distance between two nucleic acid sequences. Then, based on this distance we use the Hausdorff distance (HD) and a modified Hausdorff distance (MHD) to make the phylogenetic analysis for multi-segmented viral genomes. The results show the MHD can provide more accurate phylogenetic relationship. Our method can have global comparison of all multi-segmented genomes simultaneously, that is, we treat the multi-segmented viral genome as an entirety to make the comparative analysis. Our method is not affected by the number or order of segments, and each segment can make contribution for the phylogeny of whole genomes. We have analyzed several groups of real multi-segmented genomes from different viral families. The results show that our method will provide a new powerful tool for studying the classification of viral genomes and their phylogenetic relationships.

© 2014 Elsevier Ltd. All rights reserved.

## 1. Introduction

With the rapid development of sequencing technologies, more and more viral genome sequence information has been acquired. Characterizing genetic sequences and determining viral origins have always been important issues in virology (Holmes, 2009). It is known that the commonly used multiple sequence alignment methods fail for diverse systems of different families of RNA viruses (Holmes, 2011). Many alignment-free computational and statistical methods have been proposed for comparing viral genetic sequences (Deng et al., 2011; Pham and Zuegg, 2004; Vinga and Almeida, 2003; Yu et al., 2010; Yu et al., 2011; Yu et al., 2013). The structure variation of viral genomes is more complicated than any of those seen in the entire bacterial, plant, or

animal kingdoms. The nucleic acid comprising the viral genome may be single-stranded or double-stranded, in linear or circular configuration, and in single-segmented or multi-segmented. Multi-segmented viral genomes are those which are divided into two or more physically separate molecules of nucleic acid, all of which are then packaged into a single virus particle. For example, influenza A virus genome includes 8 single-stranded RNA segments: PB1, PB2, PA, HA, NP, NA, M1/M2, and NS1/NS2, and Glypta fumiferanae ichnovirus genome includes 105 double-stranded DNA segments. The packaging of the multi-segmented influenza virus genome into virions is a biologically intriguing event (Fujii et al., 2009). Two hypotheses have been proposed for the mechanism by which the influenza virus genome is packaged. The random packaging hypothesis suggests that each viral RNA segment possesses a common packaging signal that allows random incorporation of the RNA segments into the virions (Bancroft and Parslow, 2002). The selective-packaging hypothesis is based on the concept that each viral RNA segment possesses a unique

\* Corresponding author. Tel.: +86 10 62787874; fax: +86 10 62798033.  
E-mail address: [yau@uic.edu](mailto:yau@uic.edu) (S.-T. Yau).

packaging signal that is required for its incorporation into the virions (McGeoch et al., 1976; Odagiri and Tashiro, 1997). Therefore, in the case of unknown segment packaging signals, how to compare the multi-segmented virus genomes remains a challenging problem. Specifically, for example, given two eight-segmented influenza A virus genomes, for each one with 8 segments, we do not know exactly which segment is HA, or which segment is NA, or so on. In this case, how to compare the similarity of the two eight-segmented genomes becomes an interesting mathematical problem: how to measure the distance between two sets of eight elements. This enlightens us to develop a novel method to study the multi-segmented viral genome phylogeny.

In this paper, following Otu and Sayood work (Otu and Sayood, 2003), we use the Lempel–Ziv complexity to define the distance between two segment sequences. We classify the 42 single-segmented HIV-1 strain genomes to show the efficacy of this complexity measure. The main contribution of this work is that, based on the Lempel–Ziv complexity distance of two sequences, we use the famous Hausdorff distance (HD) and a modified Hausdorff distance (MHD) respectively to measure two multi-segmented viral genomes. The results show that MHD can provide more accurate phylogenetic relationship. Using this new method we have analyzed several groups of real multi-segmented genomes from different viral families, and find that our method is quite powerful for studying viral genome phylogeny.

## 2. Methods

### 2.1. Lempel–Ziv complexity of DNA sequence

Let  $S$  be a sequence defined over an alphabet  $\Omega$ ,  $L(S)$  be the length of  $S$ ,  $S(i)$  denotes the  $i$ th element of  $S$ , and  $S(i, j)$  defines the substring of  $S$  composed of the elements of  $S$  between positions  $i$  and  $j$  (inclusive). For DNA case,  $\Omega = \{A, C, G, T\}$ , if  $S = AACGTCGTCG$ , then  $L(S) = 10$ ,  $S(4) = G$ , and  $S(4, 7) = GTCG$ .

The Lempel–Ziv complexity of a sequence  $S$  can be measured by the minimal number of steps required for its synthesis in a certain process

$$H(S) = S(1 : i_1) \cdot S(i_1 + 1 : i_2) \cdot \dots \cdot S(i_{k-1} + 1 : i_k) \cdot \dots \cdot S(i_{m-1} + 1 : N) \cdot \quad (1)$$

At each step two operations are allowed: copying the longest fragment from the part of  $S$  which has already been synthesized, or generating a new symbol which ensures the uniqueness of each component  $S(i_{k-1} + 1 : i_k)$ .

More specifically, at each step  $k$ , the sequence  $S$  is extended by concatenating a fragment  $S(i_{k-1} + 1 : i_k)$ . The length of this fragment is 1 if some symbol at position  $i_{k-1} + 1$  occurs for the very first time. Otherwise, this fragment is obtained by copying from the prefix  $S(1 : i_{k-1})$  and adding an additional symbol. The Lempel–Ziv complexity is the number of concatenating components in this process. For example, given a DNA sequence  $S = AACGTACCATG$ , the Lempel–Ziv schema of synthesis gives the following components:  $H(S) = A \cdot \langle A \rangle \cdot C \cdot G \cdot T \cdot \langle AC \rangle \cdot C \cdot \langle A \rangle \cdot T \cdot \langle T \rangle \cdot G$  (here  $\langle \rangle$  means that the copied part from the prefix), and the corresponding complexity  $C_{LZ}(S) = 7$ . Another example is that, given a DNA sequence  $R = CTAGGGGACTTAT$ , the Lempel–Ziv schema of synthesis gives the following components

$H(R) = C \cdot T \cdot A \cdot G \cdot \langle GGG \rangle \cdot A \cdot \langle CT \rangle \cdot T \cdot \langle A \rangle \cdot T$ , and  $C_{LZ}(R) = 7$ . Note that, during one concatenating component, the part from the prefix can be continually copied many times, like  $G$  here.

Ziv and Lempel (1977) called the complexity decomposition of a sequence  $S$  following the above schema the exhaustive history of

$S$ , and mathematically proved that every sequence  $S$  has a unique exhaustive history.

### 2.2. Similarity measure by the Lempel–Ziv complexity

The Lempel–Ziv complexity provides a powerful tool for measuring the similarity between two DNA sequences (Otu and Sayood, 2003). Given two sequences  $S$  and  $R$ , consider the sequence  $SR$ , and its Lempel–Ziv complexity. By definition, the number of components needed to build  $R$  when appended to  $S$  is  $C_{LZ}(SR) - C_{LZ}(S)$ . This number will be less than or equal to  $C_{LZ}(R)$  because at any given step of the production process of  $R$  (in building the sequence  $SR$ ) we use a larger search space due to the existence of  $S$ . Therefore, if  $R$  is more similar to  $S$  than  $T$  then we would expect  $C_{LZ}(SR) - C_{LZ}(S)$  to be smaller than  $C_{LZ}(ST) - C_{LZ}(S)$ . Here we adopt a similarity measure between two sequences  $P$  and  $Q$  as

$$d(P, Q) = \frac{C_{LZ}(PQ) - C_{LZ}(P) + C_{LZ}(QP) - C_{LZ}(Q)}{\frac{1}{2}(C_{LZ}(PQ) + C_{LZ}(QP))} \quad (2)$$

because it is used and achieved a great success in the phylogenetic analysis of complete mammalian mitochondrial genomes (Otu and Sayood, 2003).

### 2.3. Comparing multi-segmented genomes

As mentioned in Section 1, measuring the distance between two multi-segmented genomes becomes a mathematical problem that measures how far two sets of multiple elements are from each other. Suppose that  $A = \{a_1, a_2, \dots, a_n\}$  and  $B = \{b_1, b_2, \dots, b_n\}$  are two  $n$ -segmented genomes, where  $a_i$  and  $b_j$  are segment sequences in genomes  $A$  and  $B$ , respectively. The distance between two segments  $a_i$  and  $b_j$  can be defined as  $d(a_i, b_j)$  by (2). The distance between a segment  $a$  and a genome  $B$  can be defined as  $d(a, B) = \min_{b \in B} d(a, b)$ . We give two distance measures between  $A$  and  $B$  below. The first one is the well-known Hausdorff distance, and the second one is a modified Hausdorff distance (Dubuisson and Jain, 1994).

Distance measure 1:

Define  $d_1(A, B) = \max_{a \in A} d(a, B)$ , then the Hausdorff distance

$$HD(A, B) = \max \{d_1(A, B), d_1(B, A)\} \quad (3)$$

Distance measure 2:

Define  $d_2(A, B) = \frac{1}{n} \sum_{a \in A} d(a, B)$ , then the modified Hausdorff distance

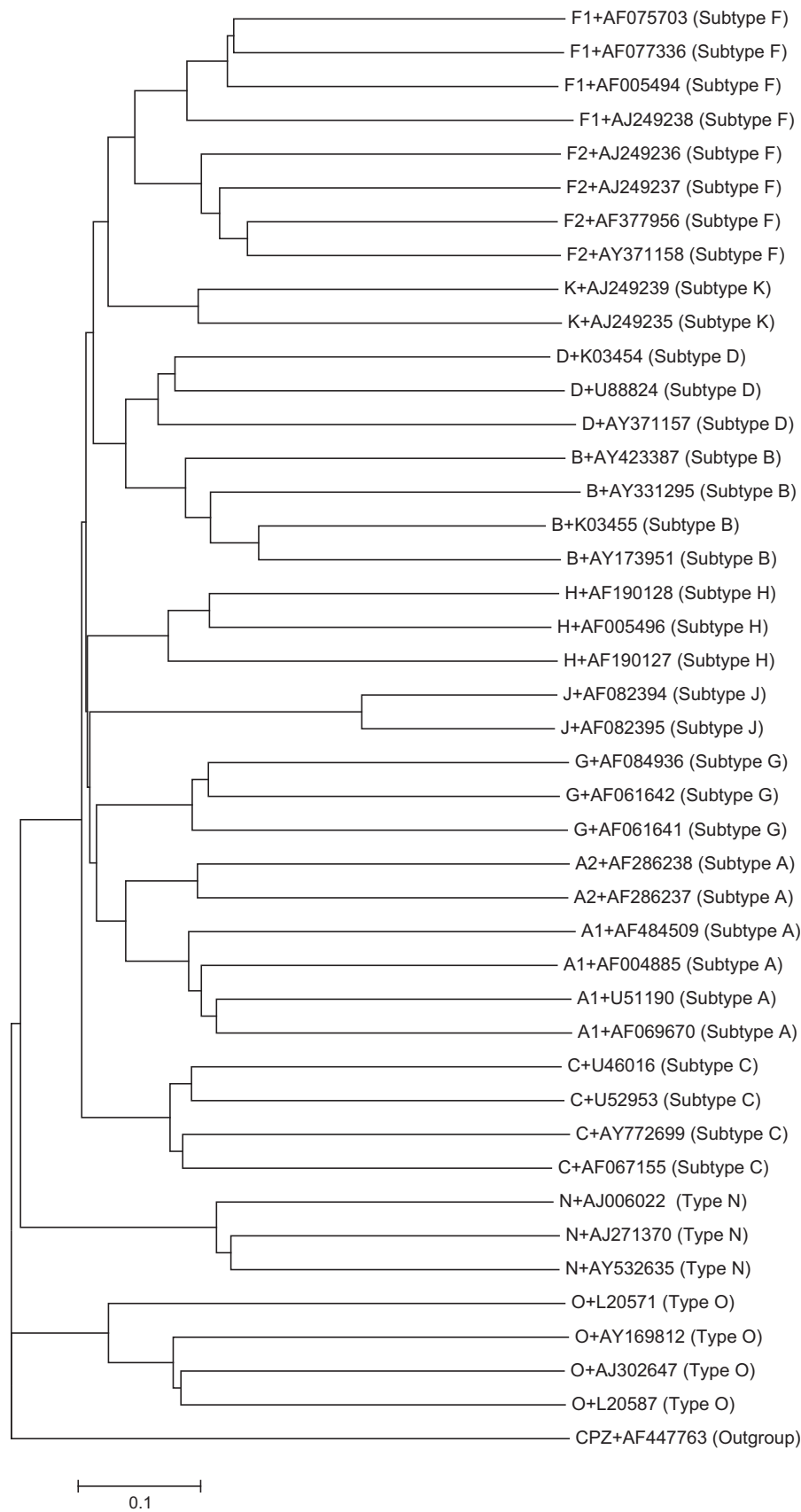
$$MHD(A, B) = \max \{d_2(A, B), d_2(B, A)\} \quad (4)$$

In the next section we show that these two distance measures can imply the phylogenetic and classification relationship between multi-segmented viral genomes.

## 3. Results and discussion

### 3.1. Phylogenetic analysis of single-segmented viral genomes

To test that the distance obtained in this way truly incorporates the classification and phylogenetic analysis of viral genomes, we firstly apply it to the real data set of single-segmented genome sequences. The 42 HIV-1 referenced sequences (Wu et al., 2007) are examined here. This data set consists of 6 subtype A (4 A1 and 2 A2), 4 subtype B, 4 subtype C, 3 subtype D, 8 subtype F (4 F1 and 4 F2), 3 subtype G, 3 subtype H, 2 subtype J, 2 subtype K, 3 type N and 4 type O. The average length of these strains is 9005 bp, with the maximum length 9829 bp and the minimum length 8349 bp. These HIV-1 reference sequences were carefully selected by considering several criteria (Leitner et al., 2005). One simian immunodeficiency virus (SIV) strain AF447763 is also added to this data set as an



**Fig. 1.** The neighbor-joining phylogenetic tree of the 42 HIV-1 strains and one SIV strain (AF447763) based on the similarity measure by the Lempel–Ziv complexity.

outgroup. We use the similarity measure by the Lempel–Ziv complexity, formula (2), to calculate the distance matrix of these 43 genomes. Then we reconstruct the phylogenetic tree of these primate

lentiviruses (Fig. 1) using a neighbor-joining algorithm (Saitou and Nei, 1987) based on MEGA 5 software (Tamura et al., 2011). In this tree, we can find that all subtypes are clearly clustered together as

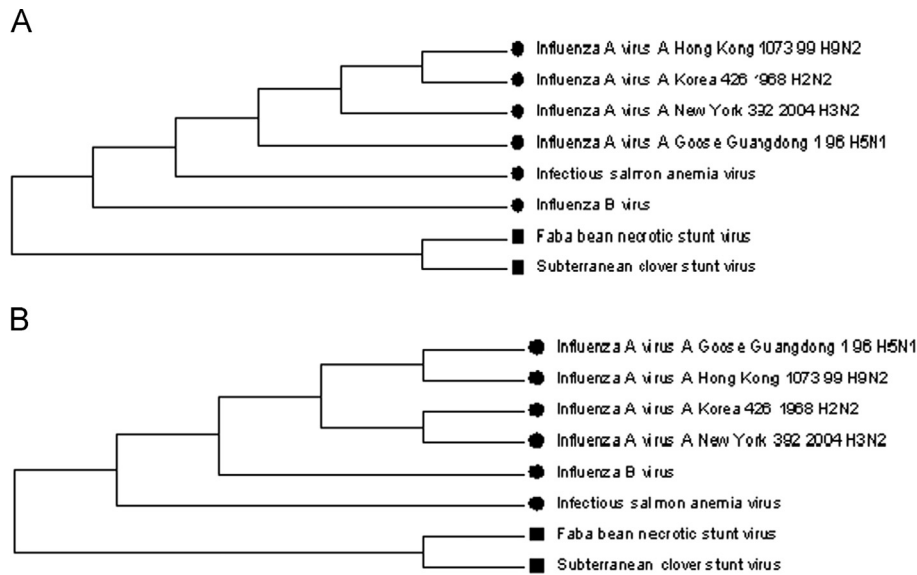
**Table 1**  
8 referenced eight-segmented viral genomes in the current GenBank collection.

No.	Virus name	Access number	Family label
1	Faba_bean_necrotic_stunt_virus_uid39929	NC_013094.1	Nanoviridae
1	Faba_bean_necrotic_stunt_virus_uid39929	NC_013095.1	Nanoviridae
1	Faba_bean_necrotic_stunt_virus_uid39929	NC_013096.1	Nanoviridae
1	Faba_bean_necrotic_stunt_virus_uid39929	NC_013097.1	Nanoviridae
1	Faba_bean_necrotic_stunt_virus_uid39929	NC_013098.1	Nanoviridae
1	Faba_bean_necrotic_stunt_virus_uid39929	NC_013099.1	Nanoviridae
1	Faba_bean_necrotic_stunt_virus_uid39929	NC_013100.1	Nanoviridae
1	Faba_bean_necrotic_stunt_virus_uid39929	NC_013101.1	Nanoviridae
2	Infectious_salmon_anemia_virus_uid15020	NC_006497.1	Orthomyxoviridae
2	Infectious_salmon_anemia_virus_uid15020	NC_006498.1	Orthomyxoviridae
2	Infectious_salmon_anemia_virus_uid15020	NC_006499.1	Orthomyxoviridae
2	Infectious_salmon_anemia_virus_uid15020	NC_006500.1	Orthomyxoviridae
2	Infectious_salmon_anemia_virus_uid15020	NC_006501.1	Orthomyxoviridae
2	Infectious_salmon_anemia_virus_uid15020	NC_006502.1	Orthomyxoviridae
2	Infectious_salmon_anemia_virus_uid15020	NC_006503.1	Orthomyxoviridae
2	Infectious_salmon_anemia_virus_uid15020	NC_006505.1	Orthomyxoviridae
3	Influenza_A_virus_A_Goose_Guangdong_1_96_H5N1__uid15617	NC_007357.1	Orthomyxoviridae
3	Influenza_A_virus_A_Goose_Guangdong_1_96_H5N1__uid15617	NC_007358.1	Orthomyxoviridae
3	Influenza_A_virus_A_Goose_Guangdong_1_96_H5N1__uid15617	NC_007359.1	Orthomyxoviridae
3	Influenza_A_virus_A_Goose_Guangdong_1_96_H5N1__uid15617	NC_007360.1	Orthomyxoviridae
3	Influenza_A_virus_A_Goose_Guangdong_1_96_H5N1__uid15617	NC_007361.1	Orthomyxoviridae
3	Influenza_A_virus_A_Goose_Guangdong_1_96_H5N1__uid15617	NC_007362.1	Orthomyxoviridae
3	Influenza_A_virus_A_Goose_Guangdong_1_96_H5N1__uid15617	NC_007363.1	Orthomyxoviridae
3	Influenza_A_virus_A_Goose_Guangdong_1_96_H5N1__uid15617	NC_007364.1	Orthomyxoviridae
4	Influenza_A_virus_A_Hong_Kong_1073_99_H9N2__uid14892	NC_004905.2	Orthomyxoviridae
4	Influenza_A_virus_A_Hong_Kong_1073_99_H9N2__uid14892	NC_004906.1	Orthomyxoviridae
4	Influenza_A_virus_A_Hong_Kong_1073_99_H9N2__uid14892	NC_004907.1	Orthomyxoviridae
4	Influenza_A_virus_A_Hong_Kong_1073_99_H9N2__uid14892	NC_004908.1	Orthomyxoviridae
4	Influenza_A_virus_A_Hong_Kong_1073_99_H9N2__uid14892	NC_004909.1	Orthomyxoviridae
4	Influenza_A_virus_A_Hong_Kong_1073_99_H9N2__uid14892	NC_004910.1	Orthomyxoviridae
4	Influenza_A_virus_A_Hong_Kong_1073_99_H9N2__uid14892	NC_004911.1	Orthomyxoviridae
4	Influenza_A_virus_A_Hong_Kong_1073_99_H9N2__uid14892	NC_004912.1	Orthomyxoviridae
5	Influenza_A_virus_A_Korea_426_1968_H2N2__uid15620	NC_007374.1	Orthomyxoviridae
5	Influenza_A_virus_A_Korea_426_1968_H2N2__uid15620	NC_007375.1	Orthomyxoviridae
5	Influenza_A_virus_A_Korea_426_1968_H2N2__uid15620	NC_007376.1	Orthomyxoviridae
5	Influenza_A_virus_A_Korea_426_1968_H2N2__uid15620	NC_007377.1	Orthomyxoviridae
5	Influenza_A_virus_A_Korea_426_1968_H2N2__uid15620	NC_007378.1	Orthomyxoviridae
5	Influenza_A_virus_A_Korea_426_1968_H2N2__uid15620	NC_007380.1	Orthomyxoviridae
5	Influenza_A_virus_A_Korea_426_1968_H2N2__uid15620	NC_007381.1	Orthomyxoviridae
5	Influenza_A_virus_A_Korea_426_1968_H2N2__uid15620	NC_007382.1	Orthomyxoviridae
6	Influenza_A_virus_A_New_York_392_2004_H3N2__uid15622	NC_007366.1	Orthomyxoviridae
6	Influenza_A_virus_A_New_York_392_2004_H3N2__uid15622	NC_007367.1	Orthomyxoviridae
6	Influenza_A_virus_A_New_York_392_2004_H3N2__uid15622	NC_007368.1	Orthomyxoviridae
6	Influenza_A_virus_A_New_York_392_2004_H3N2__uid15622	NC_007369.1	Orthomyxoviridae
6	Influenza_A_virus_A_New_York_392_2004_H3N2__uid15622	NC_007370.1	Orthomyxoviridae
6	Influenza_A_virus_A_New_York_392_2004_H3N2__uid15622	NC_007371.1	Orthomyxoviridae
6	Influenza_A_virus_A_New_York_392_2004_H3N2__uid15622	NC_007372.1	Orthomyxoviridae
6	Influenza_A_virus_A_New_York_392_2004_H3N2__uid15622	NC_007373.1	Orthomyxoviridae
7	Influenza_B_virus_uid14656	NC_002204.1	Orthomyxoviridae
7	Influenza_B_virus_uid14656	NC_002205.1	Orthomyxoviridae
7	Influenza_B_virus_uid14656	NC_002206.1	Orthomyxoviridae
7	Influenza_B_virus_uid14656	NC_002207.1	Orthomyxoviridae
7	Influenza_B_virus_uid14656	NC_002208.1	Orthomyxoviridae
7	Influenza_B_virus_uid14656	NC_002209.1	Orthomyxoviridae
7	Influenza_B_virus_uid14656	NC_002210.1	Orthomyxoviridae
7	Influenza_B_virus_uid14656	NC_002211.1	Orthomyxoviridae
8	Subterranean_clover_stunt_virus_uid14180	NC_003812.1	Nanoviridae
8	Subterranean_clover_stunt_virus_uid14180	NC_003813.1	Nanoviridae
8	Subterranean_clover_stunt_virus_uid14180	NC_003814.1	Nanoviridae
8	Subterranean_clover_stunt_virus_uid14180	NC_003815.1	Nanoviridae
8	Subterranean_clover_stunt_virus_uid14180	NC_003816.1	Nanoviridae
8	Subterranean_clover_stunt_virus_uid14180	NC_003817.1	Nanoviridae
8	Subterranean_clover_stunt_virus_uid14180	NC_003818.1	Nanoviridae
8	Subterranean_clover_stunt_virus_uid14180	NC_003819.1	Nanoviridae

distinct branches, and the closeness relationships among the subtypes are also well demonstrated. For example, the subtype A truly contains two distinguishable sub-subtypes A1 and A2, and so does the subtype F. The result shows that the similarity measure based on Lempel–Ziv complexity can successfully construct the phylogeny of single-segmented viral genomes. It is quite important for the following study of multi-segmented viral genomes.

### 3.2. Phylogenetic analysis of multi-segmented viral genomes

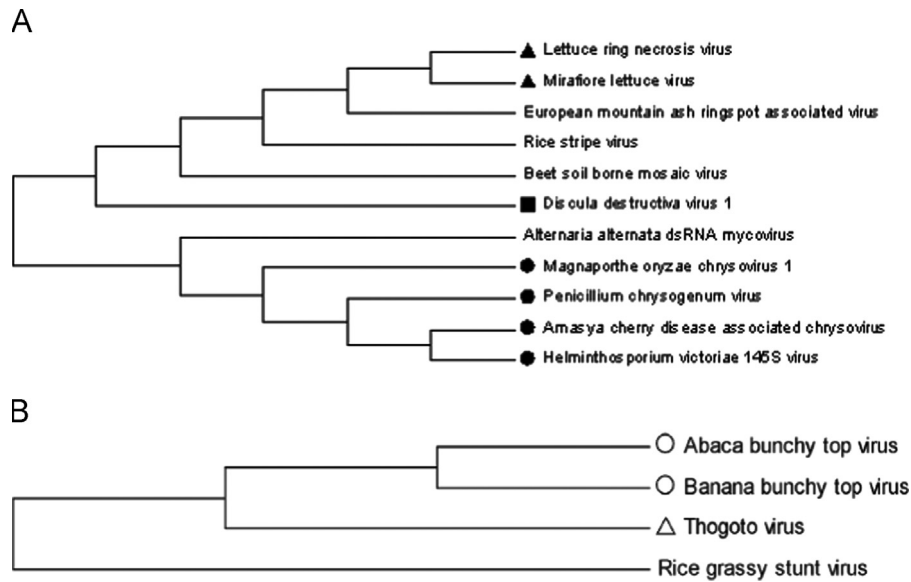
There are 8 referenced eight-segmented viral genomes in the current GenBank collection (<ftp://ftp.ncbi.nih.gov/genomes/Viruses/>). They are faba bean necrotic stunt virus, infectious salmon anemia virus, influenza A virus H5N1, influenza A virus H9N2, influenza A virus H2N2, influenza A virus H3N2, influenza B



**Fig. 2.** (A): The neighbor-joining phylogenetic tree of 8 eight-segmented referenced viral genomes based on HD. (B): The neighbor-joining phylogenetic tree of 8 eight-segmented referenced viral genomes based on MHD. Here ■ represents family *Orthomyxoviridae* and ● represents family *Nanoviridae*.

**Table 2**  
11 referenced four-segmented viral genomes in the current GenBank collection.

No.	Virus name	Access number	Family label
1	Alternaria_alternata_dsRNA_mycovirus_uid30367	NC_010984.1	Unknown
1	Alternaria_alternata_dsRNA_mycovirus_uid30367	NC_010989.1	Unknown
1	Alternaria_alternata_dsRNA_mycovirus_uid30367	NC_010990.1	Unknown
1	Alternaria_alternata_dsRNA_mycovirus_uid30367	NC_010991.1	Unknown
2	Amasya_cherry_disease_associated_chrysovirus_uid21113	NC_009944.1	Chrysoviridae
2	Amasya_cherry_disease_associated_chrysovirus_uid21113	NC_009945.1	Chrysoviridae
2	Amasya_cherry_disease_associated_chrysovirus_uid21113	NC_009946.1	Chrysoviridae
2	Amasya_cherry_disease_associated_chrysovirus_uid21113	NC_009947.1	Chrysoviridae
3	Beet_soil_borne_mosaic_virus_uid14750	NC_003503.1	Unknown
3	Beet_soil_borne_mosaic_virus_uid14750	NC_003506.1	Unknown
3	Beet_soil_borne_mosaic_virus_uid14750	NC_003507.1	Unknown
3	Beet_soil_borne_mosaic_virus_uid14750	NC_003508.1	Unknown
4	Discula_destructiva_virus_1_uid14117	NC_002797.1	Partitiviridae
4	Discula_destructiva_virus_1_uid14117	NC_002800.1	Partitiviridae
4	Discula_destructiva_virus_1_uid14117	NC_002801.1	Partitiviridae
4	Discula_destructiva_virus_1_uid14117	NC_002802.1	Partitiviridae
5	European_mountain_ash_ringspot_associated_virus_uid39973	NC_013105.1	Unknown
5	European_mountain_ash_ringspot_associated_virus_uid39973	NC_013106.1	Unknown
5	European_mountain_ash_ringspot_associated_virus_uid39973	NC_013107.1	Unknown
5	European_mountain_ash_ringspot_associated_virus_uid39973	NC_013108.1	Unknown
6	Helminthosporium_victoriae_145S_virus_uid14945	NC_005978.1	Chrysoviridae
6	Helminthosporium_victoriae_145S_virus_uid14945	NC_005979.1	Chrysoviridae
6	Helminthosporium_victoriae_145S_virus_uid14945	NC_005980.1	Chrysoviridae
6	Helminthosporium_victoriae_145S_virus_uid14945	NC_005981.1	Chrysoviridae
7	Lettuce_ring_necrosis_virus_uid14959	NC_006051.1	Ophioviridae
7	Lettuce_ring_necrosis_virus_uid14959	NC_006052.1	Ophioviridae
7	Lettuce_ring_necrosis_virus_uid14959	NC_006053.1	Ophioviridae
7	Lettuce_ring_necrosis_virus_uid14959	NC_006054.1	Ophioviridae
8	Magnaporthe_oryzae_chrysovirus_1_uid51685	NC_014462.1	Chrysoviridae
8	Magnaporthe_oryzae_chrysovirus_1_uid51685	NC_014463.1	Chrysoviridae
8	Magnaporthe_oryzae_chrysovirus_1_uid51685	NC_014464.1	Chrysoviridae
8	Magnaporthe_oryzae_chrysovirus_1_uid51685	NC_014465.1	Chrysoviridae
9	Mirafiore_lettuce_virus_uid14886	NC_004779.1	Ophioviridae
9	Mirafiore_lettuce_virus_uid14886	NC_004780.1	Ophioviridae
9	Mirafiore_lettuce_virus_uid14886	NC_004781.1	Ophioviridae
9	Mirafiore_lettuce_virus_uid14886	NC_004782.1	Ophioviridae
10	Penicillium_chrysogenum_virus_uid16141	NC_007539.1	Chrysoviridae
10	Penicillium_chrysogenum_virus_uid16141	NC_007540.1	Chrysoviridae
10	Penicillium_chrysogenum_virus_uid16141	NC_007541.1	Chrysoviridae
10	Penicillium_chrysogenum_virus_uid16141	NC_007542.1	Chrysoviridae
11	Rice_stripe_virus_uid14795	NC_003753.1	Unknown
11	Rice_stripe_virus_uid14795	NC_003754.1	Unknown
11	Rice_stripe_virus_uid14795	NC_003755.1	Unknown
11	Rice_stripe_virus_uid14795	NC_003776.1	Unknown



**Fig. 3.** (A) The neighbor-joining phylogenetic tree of 11 four-segmented referenced viral genomes based on MHD. Here ● represents family *Chrysoviridae*, ▲ represents family *Ophioviridae*, and ■ represents family *Partitiviridae*. (B) The neighbor-joining phylogenetic tree of 4 six-segmented referenced viral genomes based on MHD. Here ○ represents family *Nanoviridae*, and △ represents family *Orthomyxoviridae*.

**Table 3**

4 referenced six-segmented viral genomes in the current GenBank collection.

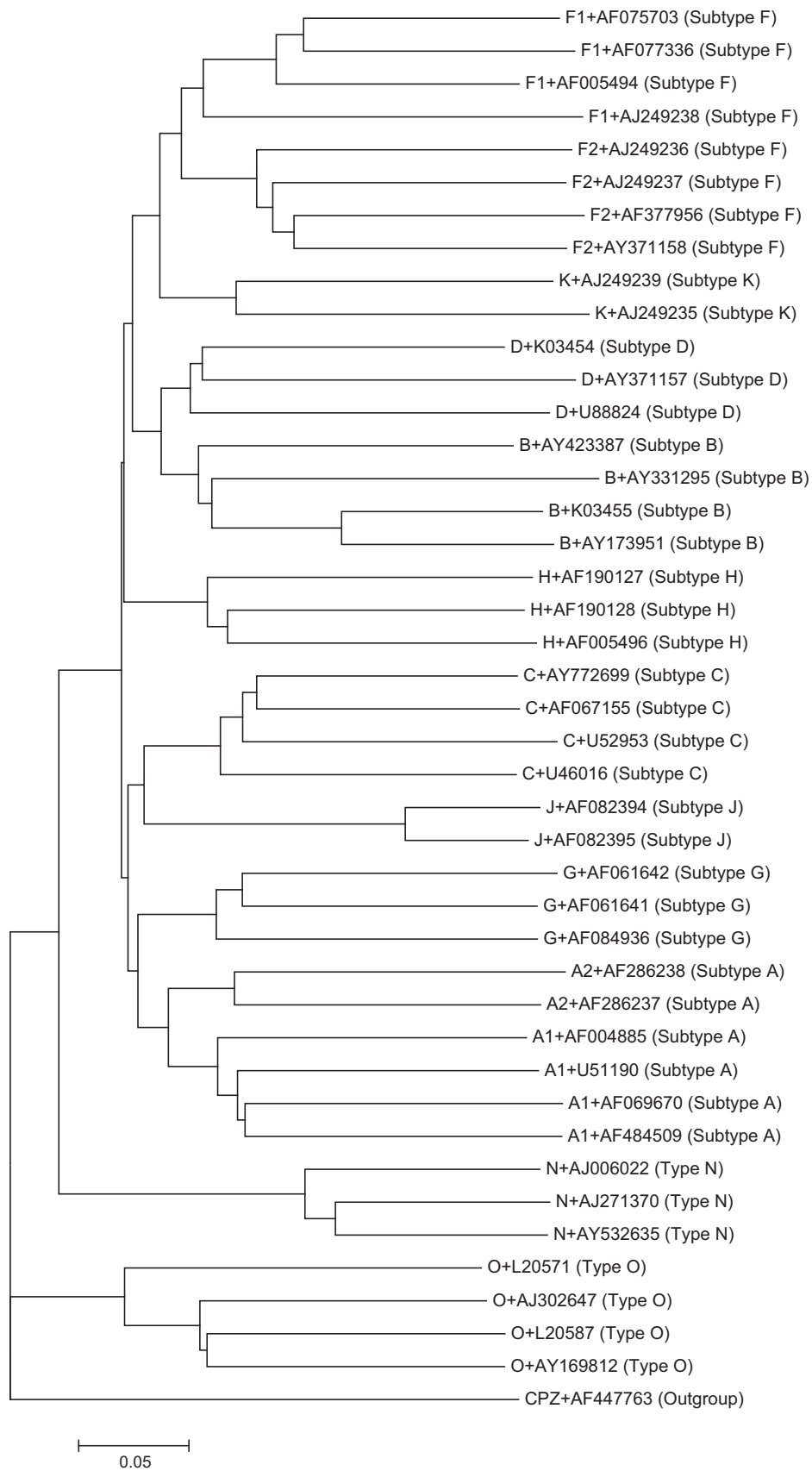
No.	Virus name	Access number	Family label
1	Abaca_bunchy_top_virus_uid28697	NC_010314.1	<i>Nanoviridae</i>
1	Abaca_bunchy_top_virus_uid28697	NC_010315.1	<i>Nanoviridae</i>
1	Abaca_bunchy_top_virus_uid28697	NC_010316.1	<i>Nanoviridae</i>
1	Abaca_bunchy_top_virus_uid28697	NC_010317.1	<i>Nanoviridae</i>
1	Abaca_bunchy_top_virus_uid28697	NC_010318.1	<i>Nanoviridae</i>
1	Abaca_bunchy_top_virus_uid28697	NC_010319.1	<i>Nanoviridae</i>
2	Banana_bunchy_top_virus_uid14621	NC_003473.1	<i>Nanoviridae</i>
2	Banana_bunchy_top_virus_uid14621	NC_003474.1	<i>Nanoviridae</i>
2	Banana_bunchy_top_virus_uid14621	NC_003475.1	<i>Nanoviridae</i>
2	Banana_bunchy_top_virus_uid14621	NC_003476.1	<i>Nanoviridae</i>
2	Banana_bunchy_top_virus_uid14621	NC_003477.1	<i>Nanoviridae</i>
2	Banana_bunchy_top_virus_uid14621	NC_003479.1	<i>Nanoviridae</i>
3	Rice_grassy_stunt_virus_uid14692	NC_002323.1	Unknown
3	Rice_grassy_stunt_virus_uid14692	NC_002324.1	Unknown
3	Rice_grassy_stunt_virus_uid14692	NC_002325.1	Unknown
3	Rice_grassy_stunt_virus_uid14692	NC_002326.1	Unknown
3	Rice_grassy_stunt_virus_uid14692	NC_002327.1	Unknown
3	Rice_grassy_stunt_virus_uid14692	NC_002328.1	Unknown
4	Thogoto_virus_uid15043	NC_006495.1	<i>Orthomyxoviridae</i>
4	Thogoto_virus_uid15043	NC_006496.1	<i>Orthomyxoviridae</i>
4	Thogoto_virus_uid15043	NC_006504.1	<i>Orthomyxoviridae</i>
4	Thogoto_virus_uid15043	NC_006506.1	<i>Orthomyxoviridae</i>
4	Thogoto_virus_uid15043	NC_006507.1	<i>Orthomyxoviridae</i>
4	Thogoto_virus_uid15043	NC_006508.1	<i>Orthomyxoviridae</i>

virus, and subterranean clover stunt virus. Each of them has eight nucleic acid segments. Furthermore, faba bean necrotic stunt virus and subterranean clover stunt virus belong to family *Orthomyxoviridae*, and the other six viruses belong to family *Nanoviridae*. For details, please see Table 1. Here we use HD and MHD to measure the distance between any two eight-segmented viral genomes, respectively. That is, by using formulae (3) and (4), we can obtain two distance matrices for the 8 viral genomes. Then we reconstruct the phylogenetic trees (see Fig. 2(A) by HD and Fig. 2(B) by MHD) of these viral genomes still using neighbor-joining algorithm based on MEGA 5 software. We find that, for both trees, faba bean necrotic stunt virus and subterranean clover stunt virus form a cluster because they are from the same family *Orthomyxoviridae*, and the other six viruses belong to another cluster (family *Nanoviridae*). Furthermore, all the influenza viruses get together

in Fig. 2(B), but Fig. 2(A) fails to show this relationship. Thus MHD can provide more accurate phylogenetic relationship for multi-segmented viral genomes. This finding is consistent with (Dubuisson and Jain, 1994) which uses MHD for object matching of synthetic images containing various levels of noise.

There are 11 referenced four-segmented viral genomes in the current GenBank collection. Each of them has four nucleic acid segments. Among these 11 viruses, amasya cherry disease associated chrysovirus, helminthosporium victoriae 145S virus, magnaporthe oryzae chrysovirus 1, and penicillium chrysogenum virus belong to family *Chrysoviridae*, lettuce ring necrosis virus and mirafiore lettuce virus belong to family *Ophioviridae*, and discula destructiva virus 1 belongs to family *Partitiviridae*. However, alternaria alternate dsRNA mycovirus, beet soil borne mosaic virus, European mountain ash ringspot associated virus, and rice stripe virus have unknown family labels so far. For details, please see Table 2. Here we use MHD to measure the distance between any two four-segmented viral genomes. That is, by using formula (4), we can obtain the distance matrix for the 11 viral genomes. Then we reconstruct the phylogenetic tree (see Fig. 3(A)) of these viral genomes using neighbor-joining algorithm based on MEGA 5 software. We find that, in the tree, the four *Chrysoviridae* family members get together, and the alternaria alternate dsRNA mycovirus is very close to this family. This finding is consistent with Aoki et al.'s work (Aoki et al., 2009) in which the authors states that this virus appears to be evolutionarily related to but not a member of the family *Chrysoviridae*. Thus we predict alternaria alternate dsRNA mycovirus belongs to a new family. The two *Ophioviridae* family members (lettuce ring necrosis virus and mirafiore lettuce virus) also get together as expected in the tree. We also find that beet soil borne mosaic virus, European mountain ash ringspot associated virus, and rice stripe virus are evolutionarily between family *Partitiviridae* and family *Ophioviridae*.

There are 4 referenced six-segmented viral genomes in the current GenBank collection. Each of them has six nucleic acid segments. Among these 4 viruses, Abaca bunchy top virus and Banana bunchy top virus belong to family *Nanoviridae*, Thogoto virus belongs to family *Orthomyxoviridae*. Rice grassy stunt virus has unknown family labels so far. For details, please see Table 3. Here we use MHD to measure the distance between any two six-segmented viral genomes. That is, by using formula (4), we can



**Fig. 4.** The neighbor-joining phylogenetic tree of the 42 HIV-1 strains and one SIV strain (AF447763) based on the composition vector method for  $K=6$ .

obtain the distance matrix for the 4 viral genomes. Then we reconstruct the phylogenetic tree (see Fig. 3(B)) of these viral genomes using neighbor-joining algorithm based on MEGA

5 software. We find that, in the tree, the two *Nanoviridae* family members get together. Rice grassy stunt virus is far away from the two families, thus we predict it belongs to a new family.

There is only one referenced member for many types of multiple-segmented viruses in the current GenBank collection. Actually, there is only 1 referenced member for 5-segmented, 7-segmented, 9-segmented, 15-segmented, 20-segmented, 24-segmented, 30-segmented, 56-segmented, and 105-segmented viruses. There is no need to reconstruct the tree for one element. For 10-segmented, 11-segmented, 12-segmented referenced viruses, they all belong to the same family *Reoviridae* or *Nanoviridae*, so the trees can not reveal more information about virus family relationship for them. Since there are a large number of referenced members for 2-segmented and 3-segmented viruses, we reconstruct the phylogenetic trees for several families of them in Section 5. In addition, there are no other multi-segmented viral genomes.

For multiple-segmented viruses, typically phylogenetic trees of each segment are reconstructed for discovering viral phylogeny. When dealing with those multi-segmented genomes, most existing methods carry out the phylogenetic analysis based on segment by segment (Lam et al., 2013). For example, the researchers usually make eight phylogenetic trees for eight segments of influenza A viruses (Lindstrom et al., 1998). These trees are usually different. Thus the phylogenetic result of whole genomes is controversial because each segment sequence generally does not contain enough information to construct an evolutionary history of organisms. On the other hand, consensus tree methods may be used to combine the phylogenetic trees based on different segments. However, consensus tree methods were not developed for instances where their segments do not match well. To the best of our knowledge, our method is the first one to globally compare viruses with multiple segments, that is, we treat the multi-segmented viral genome as an entirety to make the comparative analysis. Our method is not affected by the number or order of segments, and each segment can make contribution for the phylogeny of whole genomes. Furthermore, if we do not know the segment packaging signals as mentioned in Section 1, our method can still have global comparison of all multi-segmented genomes simultaneously which no other existing method can achieve.

The statistical tests of the resulting tree are usually used to evaluate the tree's stability. Since our approach does not use any sequence alignment, statistical re-sampling (bootstrap or jackknife) cannot be carried out in the way of random choice of nucleotide sites with replacement. Recently Zuo et al. (2010) performed time-consuming bootstrap and jackknife tests for the resulting trees obtained by the composition vector method which is also an alignment-free approach. The strategy is by randomly taking 90%, 80%, ..., 10% of proteins from the whole proteome of one species 100 times as re-sampling, and then testing the average topological distance between the 100 trees. This brings a new direction about statistical tests of the resulting tree for alignment-free methods on proteome data. Our approach uses the whole genome nucleotide sequence information not the proteome information, thus sampling of groups of proteins is not applicable. Actually, our approach does not need any parameter setting. That is, given two multi-segmented viral genomes, the distance between them by our approach is a naturally fixed value. When measuring how far two sets of multiple elements are from each other, there are many choices of distances. In the current work, we only focus on HD and MHD. Further studies in future will be needed to investigate whether other distances can bring more biologically meaningful results.

In this work, we use the Lempel–Ziv complexity to measure the distance between two nucleotide sequences. Actually, there are currently some other alignment-free tools to do it. As we mentioned above, the composition vector is also a powerful alignment-free approach (Zuo et al., 2010). However, in this  $K$ -mer-based

method,  $K$ -value controls the resolution of the whole method, i.e., different  $K$ -values will bring different results. We use  $K=4$  and  $K=5$  to make the composition vector trees based on genome nucleotide sequences of those 42 HIV-1 strains and one SIV strain. The neighbor-joining resulting trees cannot classify those subtypes together. When we use  $K=6$ , the composition vector tree (Fig. 4) shows that all subtypes can be clearly clustered together as distinct branches. Thus the composition vector method can also incorporate the classification and phylogenetic analysis of viral genomes well; however, the choice of  $K$  values depends on the data under study. On the other hand, the Lempel–Ziv complexity approach does not need any parameter setting during the process of calculating the distance.

#### 4. Conclusion

In this work, our new approach does not use multiple sequence alignment or assume any evolutionary model and it does not need this type of human intervention. The results are naturally and automatically generated. Our method can have a global comparison of all multi-segmented genomes simultaneously. The results show that our method will provide a new powerful tool for studying the classification of viral genomes and their phylogenetic relationships. The codes used to prepare this paper are available from the author upon request.

#### Acknowledgments

This research is supported by U. S. NSF grant DMS-1120824, China NSF grant 31271408, Tsinghua University start up funding, and Tsinghua University independent research project grant.

#### Appendix A. Supporting information

Supporting information associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.jtbi.2014.01.022>.

#### References

- Aoki, N., Moriyama, H., Kodama, M., Arie, T., Teraoka, T., Fukuhara, T., 2009. A novel mycovirus associated with four double-stranded RNAs affects host fungal growth in *Alternaria alternata*. *Virus Res.* 140 (1), 179–187.
- Bancroft, C.T., Parslow, T.G., 2002. Evidence for segment nonspecific packaging of the influenza A virus genome. *J. Virol.* 76, 7133–7139.
- Deng, M., Yu, C., Liang, Q., He, R.L., Yau, S.S.T., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PLoS ONE* 6 (3), e17293.
- Dubuisson, M.P., Jain, A.K., October, 1994. A modified Hausdorff distance for object matching. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, Conference A: Computer Vision & Image Processing, vol. 1 pp. 566–568.
- Fujii, K., Ozawa, M., Iwatsuki-Horimoto, K., Horimoto, T., Kawaoka, Y., 2009. Incorporation of influenza A virus genome segments does not absolutely require wildtype sequences. *J. Gen. Virol.* 90 (7), 1734–1740.
- Holmes, E.C., 2009. The comparative genomics of viral emergence. *Proc. Natl. Acad. Sci. USA* 107, 1742–1746.
- Holmes, E.C., 2011. What does virus evolution tell us about virus origins. *J. Virol.* 86, 5247–5251.
- Lam, T.T., Chong, Y.L., Shi, M., Hon, C.C., Li, J., Martin, D.P., Tang, J.W., Mok, C.K., Shih, S.R., Yip, C.W., Jiang, J., Hui, R.K., Pybus, O.G., Holmes, E.C., Leung, F.C., 2013. Systematic phylogenetic analysis of influenza A virus reveals many novel mosaic genome segments. *Infect. Genet. Evol.* 18, 367–378.
- Leitner, T., Korber, B., Daniels, M., Calef, C., & Foley, B., 2005. HIV-1 Subtype and Circulating Recombinant Form (CRF) Reference Sequences. Accessible through (<http://www.hiv.lanl.gov/content/hiv-db/REVIEWS/RefSeqs2005/RefSeqs05.html>).
- Lindstrom, S.E., Hiromoto, Y., Nerome, R., Omoe, K., Sugita, S., Yamazaki, Y., Takahashi, T., Nerome, K., 1998. Phylogenetic analysis of the entire genome of



- influenza A (H3N2) viruses from Japan: evidence for genetic reassortment of the six internal genes. *J. Virol.* 72 (10), 8021–8031.
- McGeoch, D., Fellner, P., Newton, C., 1976. Influenza virus genome consists of eight distinct RNA species. *Proc. Natl. Acad. Sci. USA* 73, 3045–3049.
- Odagiri, T., Tashiro, M., 1997. Segment-specific noncoding sequences of the influenza virus genome RNA are involved in the specific competition between defective interfering RNA and its progenitor RNA segment at the virion assembly step. *J. Virol.* 71, 2138–2145.
- Otu, H.H., Sayood, K., 2003. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics* 19 (16), 2122–2130.
- Pham, T.D., Zuegg, J., 2004. A probabilistic measure for alignment-free sequence comparison. *Bioinformatics* 20 (18), 3455–3461.
- Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4 (4), 406–425.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28 (10), 2731–2739.
- Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison—a review. *Bioinformatics* 19, 513–523.
- Wu, X., Cai, Z., Wan, X.F., Hoang, T., Goebel, R., Lin, G., 2007. Nucleotide composition string selection in HIV-1 subtyping using whole genomes. *Bioinformatics* 23 (14), 1744–1752.
- Yu, C., Deng, M., Yau, S.S.T., 2011. DNA sequence comparison by a novel probabilistic method. *Inform. Sci.* 181 (8), 1484–1492.
- Yu, C., Hernandez, T., Zheng, H., Yau, S.C., Huang, H.H., He, R.L., Yang, J., Yau, S.S.T., 2013. Real time classification of viruses in 12 dimensions. *PLoS One* 8 (5), e64328.
- Yu, C., Liang, Q., Yin, C., He, R.L., Yau, S.S.T., 2010. A novel construction of genome space with biological geometry. *DNA Res.* 17 (3), 155–168.
- Ziv, J., Lempel, A., 1977. A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory* 23 (3), 337–343.
- Zuo, G., Xu, Z., Yu, H., Hao, B., 2010. Jackknife and bootstrap tests of the composition vector trees. *Genomics Proteomics Bioinform.* 8 (4), 262–267.