Cluster analysis of the origins of the new influenza A (H1N1) virus

A Solovyov¹, G Palacios², T Briese², W I Lipkin², and R Rabadan³
R Rabadan: rabadan@dbmi.columbia.edu
¹Physics Department, Princeton University, Princeton, United States
²Center for Infection and Immunity, Mailman School of Public Health, Columbia University, New York, United States
³Department of Biomedical Informatics, Center for Computational Biology and Bioinformatics, Columbia University College of Physicians and Surgeons, New York, United States

Abstract

In March and April 2009, a new strain of influenza A(H1N1) virus has been isolated in Mexico and the United States. Since the initial reports more than 10,000 cases have been reported to the World Health Organization, all around the world. Several hundred isolates have already been sequenced and deposited in public databases. We have studied the genetics of the new strain and identified its closest relatives through a cluster analysis approach. We show that the new virus combines genetic information related to different swine influenza viruses. Segments PB2, PB1, PA, HA, NP and NS are related to swine H1N2 and H3N2 influenza viruses isolated in North America. Segments NA and M are related to swine influenza viruses isolated in Eurasia.

Introduction

Influenza A virus is a single stranded RNA virus with a segmented genome. When different influenza viruses co-infect the same cell, progeny viruses can be released that contain a novel mix of segments from both parental viruses. Since the first reported pandemic in 1918, there have been two other pandemics in the 20th century. In both cases, the pandemic strains presented a novel reassortment of genome segments derived from human and avian viruses [1–3]. The origins of the 1918 strain are so not clear, although different analyses suggest that this virus had an avian origin [4,5].

When and where pandemic reassortments happen remains a mystery. Avian viruses often undergo reassortment events among different subtypes. Several reports suggest that reassortments are also frequent between human viruses [6,7]. Swine have been found frequently with co-infections and reassortment of swine, human, and avian viruses has been reported [8–10,3]. In addition, cell surface oligosaccharide receptors of the swine trachea present both, a N-acetyleneuraminic acid-alpha2,3-galactose (NeuAcalpha2,3Gal) linkage, preferred by most avian influenza viruses, and a NeuAcalpha2,6Gal linkage, preferred by human viruses [11]. Co-infection combined with co-habitation of swine and poultry on small family farms all over Asia, and the presence of avian as well as human receptor types in pigs...
have led to the “mixing vessel” conjecture [12,13] that suggests that most of the inter-host reassortments are produced in pigs.

Recently, a new A(H1N1) subtype strain has been identified initially in Mexico, then rapidly reported in all continents. As of 27 May, 12,954 cases of the new influenza A(H1N1) virus infection, including 92 deaths have been reported to the World Health Organization [14,15]. Several approaches have been used to understand the origins of this strain. Searches in public databases containing influenza A genomes using sequence alignment tools indicated that the closest relatives for each of the eight genomic segments are from viruses circulating in swine for the past decade [16–19]. These include genome segments derived from “triple reassortant” swine viruses that combined in the late 1990s genome segments from viruses previously identified in humans, birds, and swine [20]. Similar conclusions were drawn by the application of phylogenetic techniques [16,21].

Here we present a cluster analysis using Principal Component Analysis and unsupervised clustering. Clustering methods are particularly robust under changes in the underlying evolutionary models. Our results substantiate previous reports [16,21], and demonstrate that for each of the genome segments of the new influenza A(H1N1) virus the closest relative was most recently identified in a swine, compatible with a reassortment of Eurasian and North American swine viruses (Figure 1).

**Materials and methods**

Influenza sequences were obtained from the National Center for Biotechnology Information (NCBI) [22] in the United States. We performed a search using Basic Local Alignment Search Tool (BLAST) for each of the eight A/California/04/2009(H1N1) segments separately, recording the 50 best matches. Then we constructed the union of all these matches, taking the sequences for all their segments available in the database. We aligned these sequences using the stretcher algorithm as implemented in the EMBOSS package.

After the alignment we translate the sequences into the binary data, comparing them to the reference sequence site by site. A mutation maps to 1, while a nucleotide identical to that in a reference sequence maps to 0. Whenever there are masks, they map to the corresponding fractional numbers. Gaps are not counted as polymorphisms. Therefore, if there are the S sequences restricted to the P polymorphic sites, these data translate to the $S \times P$ matrix. Each row of this matrix can be thought of as a vector in a P-dimensional space, and it represents one of the sequences.

We perform the Principal Component Analysis (PCA) in order to determine the most significant coordinates in this P-dimensional space. After this we leave the principal components which capture 85% of the total variance, discard the remaining ones and project the data onto this relevant coordinate subset.

This procedure is followed by the consensus K-means clustering. Namely, if one targets for K clusters, one repeats the K-means clustering procedure N times, and forms the matrix n whose elements $n_{ij} (i,j=1,\ldots,S)$ represent the number of times out of the N trials when the i-
th and j-th sequences were clustered together. In our analysis we set $N \geq 100$. The matrix of the distances between the samples is:

$$D_{ij} = 1 - \frac{n_{ij}}{N}.$$  

One then performs the standard hierarchical clustering with this matrix, targeting for the $K$ clusters. This procedure does not depend on any assumptions made by the phylogenetic models. Note that these techniques can be used for inferring phylogenies as well [23], though this is beyond the scope of the present note.

**Results**

Sequence comparison of available sequences of the new A(H1N1) virus (as of 27 May 2009) did not identify significant sequence variation, except for a few point mutations. Hence A/California/04/2009(H1N1) was chosen as the representative for further analyses.

There are many different phylogenetic techniques, each of them with their own assumptions about evolutionary models that vary in the way of computing genetic distances, probabilities, etc. As opposed to phylogenetic techniques, cluster methods do not have a need for evaluation of a tree, which is a more complicated structure than a set of clusters. Clustering techniques do not provide a detailed phylogenetic structure because they analyse group features of the sequence data. That is why the clustering analysis is more robust to the assumptions we make, for instance, the choice of genetic distance. Unsupervised methods provide a way of identifying clusters without relying on previous information about the origins, host and time isolation.

Figures 2a–2h show the data projected onto the first two principal components with the corresponding percentage of variation. The figures clearly show that in all cases the new virus sequences clustered with those of swine viruses. The closest matches for each of the segments are summarised in the Table.

Our analyses support the hypotheses whereby the 2009 pandemic influenza A(H1N1) virus derives from one or multiple reassortment(s) between influenza A viruses circulating in swine in Eurasia and in North America. It is schematically illustrated in the Figure 1.

Supplementary Tables 1 to 8 show the results of the clustering for each of the eight segments (PB2, PB1, PA, HA, NP, NA, M NS):

- [http://www.eurosurveillance.org/public/public_pdf/Table_1_Cluster_analysis_HA.pdf](http://www.eurosurveillance.org/public/public_pdf/Table_1_Cluster_analysis_HA.pdf)
- [http://www.eurosurveillance.org/public/public_pdf/Table_2_Cluster_analysis_NA.pdf](http://www.eurosurveillance.org/public/public_pdf/Table_2_Cluster_analysis_NA.pdf)
- [http://www.eurosurveillance.org/public/public_pdf/Table_3_Cluster_analysis_PB2.pdf](http://www.eurosurveillance.org/public/public_pdf/Table_3_Cluster_analysis_PB2.pdf)
- [http://www.eurosurveillance.org/public/public_pdf/Table_4_Cluster_analysis_PB1.pdf](http://www.eurosurveillance.org/public/public_pdf/Table_4_Cluster_analysis_PB1.pdf)
- [http://www.eurosurveillance.org/public/public_pdf/Table_5_Cluster_analysis_PA.pdf](http://www.eurosurveillance.org/public/public_pdf/Table_5_Cluster_analysis_PA.pdf)
- [http://www.eurosurveillance.org/public/public_pdf/Table_6_Cluster_analysis_NP.pdf](http://www.eurosurveillance.org/public/public_pdf/Table_6_Cluster_analysis_NP.pdf)
Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The work of T. Briese, G. Palacios and W. I. Lipkin was supported by National Institutes of Health awards HL083850 and AI57158 (Northeast Biodefense Center - Lipkin). The work of A. Solovyov has been supported by grant NSF PHY-0756966.

References

1. Webster RG, Laver WG. Studies on the origin of pandemic influenza. I. Antigenic analysis of A 2 influenza viruses isolated before and after the appearance of Hong Kong influenza using antisera to the isolated hemagglutinin subunits. Virology. 1972; 48(2):433–444. [PubMed: 4112502]
2. Kawaoka Y, Krauss S, Webster RG. Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. J Virol. 1989; 63(11):4603–4608. [PubMed: 2795713]
3. Scholtissek C, von Hoyningen V, Rott R. Genetic relatedness between the new 1977 epidemic strains (H1N1) of influenza and human influenza strains isolated between 1947 and 1957 (H1N1). Virology. 1978; 89(2):613–617. [PubMed: 716220]
4. Taubenberger JK, Reid AH, Lourens RM, Wang R, Jin G, Fanning TG. Characterization of the 1918 influenza virus polymerase genes. Nature. 2005; 437(7060):889–93. [PubMed: 16208372]
5. Rabdan R, Levine AJ, Robins H. Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes. J Virol. 2006 Dec; 80(23):11887–91. [PubMed: 16987977]
6. Rabdan R, Levine AJ, Krasnitz M. Non-random reassortment in human influenza A viruses. Influenza Other Respi Viruses. 2008; 2(1):9–22.
7. Nelson MI, Viboud C, Simonsen L, Bennett RT, Griesemer SB, St George K, et al. Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. PLoS Pathog. 2008 Feb 29;4(2):e1000012. [PubMed: 18463694]
8. Zhou NN, Senne DA, Landgraf JS, Swenson SL, Erickson G, Rossow K, et al. Genetic reassortment of avian, swine, and human influenza A viruses in American pigs. J Virol. 1999; 73(10):8851–6. [PubMed: 10482643]
9. Webby RJ, Swenson SL, Krauss SL, Gerrish PJ, Goyal SM, Webster RG. Evolution of swine H3N2 influenza viruses in the United States. J Virol. 2000; 74(18):8243–51. [PubMed: 10954521]
10. Lindstrom SE, Cox NJ, Klimov A. Genetic analysis of human H2N2 and early H3N2 influenza viruses, 1957–1972: evidence for genetic divergence and multiple reassortment events. Virology. 2004; 328(1):101–19. [PubMed: 15380362]
11. Ito T, Couceiro JN, Kelm S, Baum LG, Krauss S, Castrucci MR, et al. Molecular basis for the generation in pigs of influenza A viruses with pandemic potential. J Virol. 1998; 72(9):7367–73. [PubMed: 9696833]
12. Scholtissek C. Pigs as the ‘mixing vessel’ for the creation of new pandemic influenza A viruses. Med Princip Prac. 1990; 2:65–71.
13. vanReeth K. Avian influenza in swine: a threat for the human population. Verh K Acad Geneeskd Belg. 2006; 68(2):81–101. [PubMed: 16800240]
14. World Health Organization (WHO). Influenza A(H1N1). Available from: http://www.who.int/csr/disease/swineflu/en/index.html
15. Centers for Disease Control and Prevention (CDC). H1N1 Swine flu. Available from: http://www.cdc.gov/h1n1flu/
16. Trifonov V, Khiabanian H, Greenbaum B, Rabadan R. The origin of the recent swine influenza A(H1N1) virus infecting humans. Euro Surveill. 2009; 14(17) pii=19193. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19193.

17. Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, et al. Antigenic and Genetic Characteristics of Swine-Origin 2009 A(H1N1) Influenza Viruses Circulating in Humans. Science. May 22.2009 Epub ahead of print. 10.1126/science.1176225

18. Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team. Emergence of a Novel Swine-Origin Influenza A (H1N1) Virus in Humans. N Engl J Med. May 22.2009 Epub ahead of print.

19. Trifonov V, Khiabanian H, Rabadan R. Geographic Dependence, Surveillance, and Origins of the 2009 Influenza A (H1N1) Virus. New England Journal of Medicine, NEJM. May 27.2009 Epub ahead of print. 10.1056/NEJMp0904572

20. Shinde V, Bridges CB, Uyeki TM, Shu B, Balish A, Xu X, et al. Triple-Reassortant Swine Influenza A (H1) in Humans in the United States, 2005–2009. N Engl J Med. May 22.2009 Epub ahead of print.

21. Rambaut, A. Human/Swine A/H1N1 Influenza Origins and Evolution. May 3. 2009 Available from: http://tree.bio.ed.ac.uk/groups/influenza/

22. National Center for Biotechnology Information. Influenza virus resource, information, search and analysis. Available from: http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html

23. Alexe G, Satya RV, Seiler M, Platt D, Bhanot T, Hui S, et al. PCA and clustering reveal alternate mtDNA phylogeny of N and M clades. J Mol Evol. 2008; 67(5):465–87. [PubMed: 18855041]
Figure 1. Origins of the new influenza A(H1N1) virus
Schematic representation of the main results of the cluster analysis. The analysis shows that the recent A(H1N1) virus is a reassortment of at least two swine influenza viruses from North America (in light blue) and Eurasia (in dark blue).
The diagram shows a scatter plot with the first principal component on the x-axis, ranging from -15 to 15, and the second principal component on the y-axis, ranging from -12 to 4. The plot indicates different clusters of data points:

- Swine H3, Eurasia
- Human and Swine/Avian H3, North America
- Swine H1, North America
- Swine H1 and Avian H1'5'6'7, Eurasia
- Recent H1N1 virus

The graph suggests a high variance in the data, with 50% of the total variance explained by the first principal component and 8% explained by the second principal component.
First principal component, 38% of the total variance

Second principal component, 8% of the total variance

- Swine N1, North America
- Swine/Avian N1, Eurasia
- Swine N2, Eurasia
- Human and Swine N2, North America
- Swine and Human N1', Eurasia
- Recent H1N1 virus
Figure 2.
Figure 2a. Cluster analysis of the new influenza A(H1N1) virus. HA segment; data projected onto the first two principal components
Representation of the segment by segment analysis of the closest relatives to the 2009 H1N1 influenza viruses. Data is projected in the two axes of maximal variation. Clusters in different colors represent a distinct host and geographic location. Segments PB2, PB1, PA, HA, NP and NS are related to viruses isolated in swine in North America and NA and MP to swine viruses in Eurasia

Figure 2b. Cluster analysis of the new influenza A(H1N1) virus. NA segment; data projected onto the first two principal components
Representation of the segment by segment analysis of the closest relatives to the 2009 H1N1 influenza viruses. Data is projected in the two axes of maximal variation. Clusters in different colors represent a distinct host and geographic location. Segments PB2, PB1, PA, HA, NP and NS are related to viruses isolated in swine in North America and NA and MP to swine viruses in Eurasia

Figure 2c. Cluster analysis of the new influenza A(H1N1) virus. PB2 segment; data projected onto the first two principal components
Representation of the segment by segment analysis of the closest relatives to the 2009 H1N1 influenza viruses. Data is projected in the two axes of maximal variation. Clusters in different colors represent a distinct host and geographic location. Segments PB2, PB1, PA, HA, NP and NS are related to viruses isolated in swine in North America and NA and MP to swine viruses in Eurasia.
different colors represent a distinct host and geographic location. Segments PB2, PB1, PA, HA, NP and NS are related to viruses isolated in swine in North America and NA and MP to swine viruses in Eurasia.

Figure 2d. Cluster analysis of the new influenza A(H1N1) virus. PB1 segment; data projected onto the first two principal components

Representation of the segment by segment analysis of the closest relatives to the 2009 H1N1 influenza viruses. Data is projected in the two axes of maximal variation. Clusters in different colors represent a distinct host and geographic location. Segments PB2, PB1, PA, HA, NP and NS are related to viruses isolated in swine in North America and NA and MP to swine viruses in Eurasia.

Figure 2e. Cluster analysis of the new influenza A(H1N1) virus. PA segment; data projected onto the first two principal components

Representation of the segment by segment analysis of the closest relatives to the 2009 H1N1 influenza viruses. Data is projected in the two axes of maximal variation. Clusters in different colors represent a distinct host and geographic location. Segments PB2, PB1, PA, HA, NP and NS are related to viruses isolated in swine in North America and NA and MP to swine viruses in Eurasia.

Figure 2f. Cluster analysis of the new influenza A(H1N1) virus. NP segment; data projected onto the first two principal components

Representation of the segment by segment analysis of the closest relatives to the 2009 H1N1 influenza viruses. Data is projected in the two axes of maximal variation. Clusters in different colors represent a distinct host and geographic location. Segments PB2, PB1, PA, HA, NP and NS are related to viruses isolated in swine in North America and NA and MP to swine viruses in Eurasia.

Figure 2g. Cluster analysis of the new influenza A(H1N1) virus. MP segment; data projected onto the first two principal components

Representation of the segment by segment analysis of the closest relatives to the 2009 H1N1 influenza viruses. Data is projected in the two axes of maximal variation. Clusters in different colors represent a distinct host and geographic location. Segments PB2, PB1, PA, HA, NP and NS are related to viruses isolated in swine in North America and NA and MP to swine viruses in Eurasia.

Figure 2h. Cluster analysis of the new influenza A(H1N1) virus. NS segment; data projected onto the first two principal components

Representation of the segment by segment analysis of the closest relatives to the 2009 H1N1 influenza viruses. Data is projected in the two axes of maximal variation. Clusters in different colors represent a distinct host and geographic location. Segments PB2, PB1, PA, HA, NP and NS are related to viruses isolated in swine in North America and NA and MP to swine viruses in Eurasia.
Table

Closer clusters to the new influenza A(H1N1) virus.

| Segment | Closest match       | Years     |
|---------|---------------------|-----------|
| PB2     | Swine, North America| 1998–2007 |
| PB1     | Swine, North America| 1998–2007 |
| PA      | Swine, North America| 1998–2007 |
| HA      | Swine H1, North America| 1985–2007 |
| NP      | Swine, North America| 1985–2007 |
| NA      | Avian/Swine N1, Eurasia| 1982–2007 |
| M       | Swine, Eurasia      | 1980–2005 |
| NS      | Swine, North America| 1998–2007 |

Closer clusters to each of the segments of the new influenza A(H1N1) virus. The analysis reveals two clusters of related viruses: North American swine viruses (in light blue) and Eurasian swine viruses (in dark blue).