Reassortment of Ancient Neuraminidase and Recent Hemagglutinin in Pandemic (H1N1) 2009 Virus

Priyasma Bhoumik and Austin L. Hughes

Sequence analyses show that the outbreak of pandemic (H1N1) 2009 resulted from the spread of a recently derived hemagglutinin through a population of ancient and more diverse neuraminidase segments. This pattern implies reassortment and suggests that the novel form of hemagglutinin conferred a selective advantage.

Influenza virus A is a single-strand, negative-sense RNA virus whose genome consists of 8 RNA segments that encode 10 proteins (*1*). Influenza A is endemic in wild waterfowl, from which new strains periodically emerge to infect mammals, including humans and domestic pigs (*2*). Strains of influenza A viruses are categorized according to serotypes for hemagglutinin (HA) and neuraminidase (NA) proteins. These proteins cover the surface of the virus, are the main targets of the host's cellular immune response, and play major roles in the infection process (*1*,*3*,*4*).

In 2009, a novel strain of influenza A virus, pandemic (H1N1) 2009 virus, appeared in the human population, infecting thousands and causing many deaths (*2*,*5*–*8*). Phylogenetic analyses support a close relationship between the new strain and the strains that infect swine (*6*–*9*). Because different segments of the pandemic (H1N1) 2009 virus genome show different patterns of relationship to previously identified clades of influenza A virus sequences, these analyses support a role for intersegment reassortment in the origin of the new strain (*6*–*9*). For example, HA of pandemic (H1N1) 2009 virus shows a close relationship to that of classical swine influenza A virus, and NA shows a close relationship to that of Eurasian swine influenza A virus (*6*–*9*).

The Study

To examine the effects of intersegment reassortment on sequence diversity, we analyzed the pattern of nucleotide substitutions in pandemic (H1N1) 2009 virus and compared it with that of other influenza A virus genotypes (see www.biol.sc.edu/~austin). In pandemic (H1N1) 2009 virus, synonymous (π_S) and nonsynonymous (π_N) nucleotide diversity (online Technical Appendix, www.cdc.gov/EID/content/16/11/1748-Techapp.pdf) was significantly greater in NA than in HA (Table 1). In pandemic (H1N1) 2009 virus, π_S in NA was >100× that in HA, and π_N in NA was >50× times that in HA (Table 1). By contrast, in pre-2009 influenza virus subtype H1N1, π_S and π_N were similar in HA and NA (Table 1). Likewise, in influenza virus subtypes H3N2 and H5N1, π_S and π_N were similar in HA and NA (Table 1). Thus, pandemic (H1N1) 2009 virus was unique among serotypes in showing a marked difference in sequence diversity between HA and NA.

To test whether the difference between HA and NA in pandemic (H1N1) 2009 virus resulted from sampling error, we applied the same analysis to 92 epidemiologically matched pairs of HA and NA sequences from pandemic (H1N1) 2009 virus (see www.biol.sc.edu/~austin) collected in the same month (the same date, when possible) and from the same state (or the same country if not of US origin). π_S was significantly greater in NA (mean ± SE 0.2537 ± 0.0183) than in HA (0.0030 ± 0.0011; $p<0.001$ by z-test). Likewise, in epidemiologically matched pairs, π_N was significantly greater in NA (0.0215 ± 0.0022) than in HA (0.0012 ± 0.0003; $p<0.001$ by z-test).

In HA and NA genes of serotypes of influenza subtypes H1N1 (pre-2009), H3N2, and H5N1, π_S was significantly greater than π_N (Table 1). For pandemic (H1N1) 2009, π_S was significantly greater than π_N in NA (Table 1); π_S was also greater than π_N in HA, but the difference was not significant because diversity was low at synonymous and nonsynonymous sites (Table 1). π_S was significantly greater than π_N for each of the other 6 genes (online Technical Appendix Table). A pattern of π_S greater than π_N indicates past purifying selection that has eliminated deleterious nonsynonymous mutations (*10*).

To obtain evidence regarding slightly deleterious variants subject to ongoing purifying selection (*11*–*13*), we examined gene diversity at synonymous and nonsynonymous polymorphic single-nucleotide polymorphism (SNP) sites in HA and NA genes (Table 2). In the NA genes of pandemic (H1N1) 2009 virus, subtypes H1N1 (pre-2009), H3N2, and H5N1, the gene diversity at nonsynonymous SNP sites was significantly lower than that at synonymous SNP sites (Table 2). The same pattern was seen in SNP sites in the HA gene of all serotypes except pandemic (H1N1) 2009 virus. Thus, the HA gene of pandemic (H1N1) 2009 virus showed a unique pattern in the absence of evidence of ongoing purifying selection decreasing the frequency of slightly deleterious variants.

Author affiliation: University of South Carolina, Columbia, South Carolina, USA

DOI: 10.3201/eid1611.100361

Table 1. Synonymous and nonsynonymous nucleotide diversity in hemagglutinin and neuraminidase genes of influenza A virus genotypes*

	HA			NA		
Genotype	No. sequences	π_S ± SE	π_N ± SE	No. sequences	π_S ± SE	π_N ± SE
Pandemic H1N1 (2009)	397	0.0041 ± 0.0015	0.0012 ± 0.0003	171	0.4626 ± 0.0493†	0.0616 ± 0.0065†
H1N1 (pre-2009)	105	0.0926 ± 0.0063	0.0171 ± 0.0017	105	0.0842 ± 0.0088	0.0126 ± 0.0016
H3N2	562	0.1178 ± 0.0094	0.0229 ± 0.0028	357	0.0871 ± 0.0077	0.0213 ± 0.0020
H5N1	109	0.0918 ± 0.0080	0.0189 ± 0.0026	116	0.1034 ± 0.0082	0.0194 ± 0.0027

*HA, hemagglutinin; NA, neuraminidase; π_S, synonymous nucleotide diversity; π_N, nonsynonymous nucleotide diversity. There was a significant difference ($p<0.001$) between π_S and π_N in all cases except HA of pandemic (H1N1) 2009.
†Significant difference between π_S or π_N in NA and corresponding value in HA (z-test; $p<0.001$).

Table 2. Mean ± SE gene diversity at synonymous and nonsynonymous polymorphic nucleotide sites in hemagglutinin and neuraminidase genes of influenza A virus serotypes*

	HA		NA	
Genotype	Synonymous	Nonsynonymous	Synonymous	Nonsynonymous
Pandemic H1N1 (2009)	0.0120 ± 0.0007 [173]	0.0112 ± 0.0006 [839]	0.2535 ± 0.0173 [179]	0.0863 ± 0.0060 [706]†
H1N1 (pre-2009)	0.0798 ± 0.0063 [198]	0.0506 ± 0.0019 [814]†	0.0765 ± 0.0083 [152]	0.0453 ± 0.0017 [712]†
H3N2	0.0710 ± 0.0086 [203]	0.0331 ± 0.0030 [793]†	0.0760 ± 0.0087 [177]	0.0332 ± 0.0029 [688]†
H5N1	0.1195 ± 0.0106 [184]	0.0552 ± 0.0027 [834]†	0.1120 ± 0.0098 [157]	0.0482 ± 0.0024 [645]†

*HA, hemagglutinin; NA, neuraminidase. Numbers of polymorphic nucleotide sites are indicated in brackets.
†Gene diversity at nonsynonymous sites significantly different from that at synonymous sites ($p<0.001$; randomization test).

At 9 aa positions in HA, a residue not seen in our sample of pre-2009 influenza (H1N1) virus was fixed (100% frequency) in our sample of pandemic (H1N1) 2009 virus (Figure). The following amino acid replacements were involved; residue(s) in pre-2009 influenza (H1N1) are listed first: F/I/L88S, N101S, T256K, N/S275E, A/D/G277N, Q382L, G/R391E, F454Y, and S510A. Of these positions, 4 (88, 101, 275, and 391) were among those listed as having unique amino acid residues in pandemic (H1N1) 2009 virus on the basis of a smaller sequence sample by Ding et al. (*9*).

Conclusions

Analysis of nucleotide sequences of HA and NA from 4 serotypes of influenza A virus showed a unique pattern of polymorphism in pandemic (H1N1) 2009 virus. In other serotypes, diversity of synonymous and nonsynonymous nucleotides was similar in HA and NA; in pandemic (H1N1) 2009 virus, HA showed much lower nucleotide diversity at synonymous and nonsynonymous sites than did NA. Of all serotypes analyzed, NA showed evidence of past and ongoing purifying selection against deleterious nonsynonymous mutations, and HA showed evidence of past and ongoing purifying selection of all serotypes except pandemic (H1N1) 2009 virus. These unique features of HA of pandemic (H1N1) 2009 virus imply that it has a more recent common ancestor than NA of the same serotype and that it has spread rapidly by frequent reassortment into a background of a much more ancient NA genotype.

The recent spread of HA of pandemic (H1N1) 2009 virus implies multiple events of reassortment, creating a population of viruses with an ancient and diverse NA gene and a much less diverse HA gene. The polymerase basic protein 1 gene also showed low diversity (online Technical Appendix Table), suggesting similar reassortment. Other genes of pandemic (H1N1) 2009 virus showed a level of diversity intermediate between that of HA and NA, suggesting that their association with this ancient and diverse NA may have resulted from earlier reassortment events. The bottleneck in the history of HA of pandemic (H1N1) 2009 virus explains the low genetic diversity and the absence of evidence of ongoing purifying selection because purifying selection is most effective when the population is large (*11–13*). Absence of ongoing purifying selection is thus consistent with a recent population expansion, of which pandemic (H1N1) 2009 virus shows evidence (*14*).

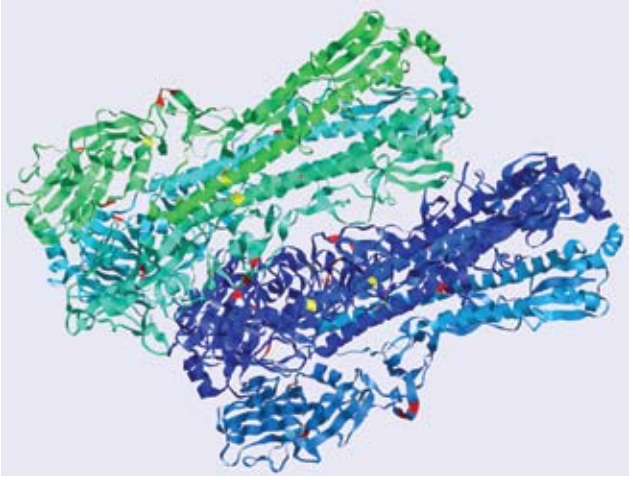

Figure. Structure of the pandemic (H1N1) 2009 virus hemagglutinin homotrimer, indicating (in red) the 9 aa positions in hemagglutinin at which a residue not found in pre-2009 influenza (H1N1) was fixed (100% frequency) in pandemic (H1N1) 2009 virus.

One factor that might have favored the spread of a recently evolved HA segment in the pandemic (H1N1) 2009 virus population would be the occurrence of ≥1 selectively favored aa replacements, causing a selective sweep (*15*) and reducing diversity at the HA locus. Such replacements in the ancestor of pandemic (H1N1) 2009 virus would likely be conserved in the pandemic (H1N1) 2009 virus population. The 9 aa residues in HA not found in our sample of pre-2009 influenza (H1N1), but fixed in our sample of pandemic (H1N1) 2009 virus, are candidates for selectively favored amino acid replacements in pandemic (H1N1) 2009 virus. Low diversity in ≥1 genes may be a recurring feature of newly emerged influenza A pandemics, supporting the need for vaccine development early in a pandemic to minimize mutation accumulation in viral genes of low initial variability.

This research was supported by grant GM43940 from the National Institutes of Health to A.L.H.

Dr Bhoumik recently completed her PhD degree at the University of South Carolina, working on the molecular evolution of viruses.

Dr Hughes is a Carolina Distinguished Professor in the Department of Biological Sciences at the University of South Carolina. His research focuses on the population genetics and molecular evolution of the immune system and of major pathogens, including viruses and malaria parasites, of humans and other vertebrates.

References

1. Brown EG. Influenza virus genetics. Biomed Pharmacother. 2000;54:169–2009. DOI: 10.1016/S0753-3322(00)89026-5
2. Gatherer D. The 2009 H1N1 influenza outbreak in its historical context. J Clin Virol. 2009;45:174–8. DOI: 10.1016/j.jcv.2009.06.004
3. Colman PM. Influenza virus neuraminidase: structure, antibodies, and inhibitors. Protein Sci. 1994;3:1687–96. DOI: 10.1002/pro.5560031007
4. Wiley DC, Skehel JJ. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. Annu Rev Biochem. 1987;56:365–94. DOI: 10.1146/annurev.bi.56.070187.002053
5. Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team, Dawood FS, Jain S, Finelli L, Shaw MS, Lindstrom S, Garten RJ, et al. Emergence of a novel swine-origin influenza A (H1N1) virus in humans. N Engl J Med. 2009;360:2605–15.
6. 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. 2009;325:197–201. DOI: 10.1126/science.1176225
7. Peiris JS, Poon LL, Guan Y. Emergence of a novel swine-origin influenza A virus (S-OIV) H1N1 virus in humans. J Clin Virol. 2009;45:169–73. DOI: 10.1016/j.jcv.2009.06.006
8. Schnitzler SU, Schnitzler P. An update on swine-origin influenza A/H1N1: a review. Virus Genes. 2009;39:279–92. DOI: 10.1007/s11262-009-0404-8
9. Ding N, Wu N, Xu Q, Chen K, Zhang C. Molecular evolution of novel swine-origin A/H1N1 influenza viruses among and before human. Virus Genes. 2009;39:293–300. DOI: 10.1007/s11262-009-0393-7
10. Hughes AL. Adaptive evolution of genes and genomes. New York: Oxford University Press; 1999.
11. Hughes AL. Near neutrality: leading edge of the neutral theory of molecular evolution. Ann N Y Acad Sci. 2008;1133:162–79. DOI: 10.1196/annals.1438.001
12. Hughes AL. Small effective population sizes and rare nonsynonymous variants in potyviruses. Virology. 2009;393:127–34. DOI: 10.1016/j.virol.2009.07.016
13. Hughes AL, Packer B, Welsch R, Bergen AW, Chanock SJ, Yeager M. Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc Natl Acad Sci U S A. 2003;100:15754–7. DOI: 10.1073/pnas.2536718100
14. Goñi N, Fajardo A, Moratorio G, Colina R, Cristina J. Modeling gene sequences over time in 2009 H1N1 influenza A virus populations. Virol J. 2009;6:215. DOI: 10.1186/1743-422X-6-215
15. Maynard Smith J, Haigh J. The hitch-hiking effect of a favorable gene. Genet Res. 1974;23:23–35. DOI: 10.1017/S0016672300014634

Address for correspondence: Austin L. Hughes, Department of Biological Sciences, Coker Life Sciences Bldg, University of South Carolina, 700 Sumter St, Columbia, SC 29208, USA; email: austin@biol.sc.edu

Article DOI: 10.3201/eid1611.100361

Reassortment of Ancient Neuraminidase and Recent Hemagglutinin in Pandemic (H1N1) 2009 Virus

Technical Appendix

Supplementary Methods

Sequences

We downloaded the coding sequences from 4 serotypes of influenza virus—H5N1, H3N2, H1N1 (pre-2009) and H1N1 (2009) from the NCBI Influenza Virus Resource (www.ncbi.nlm.nih.gov/genomes/FLU/Database/select.cgi?go=1). In the case of H1N1 (2009), we used only sequences deposited between January and July 2009 to obtain a representative picture of worldwide diversity at the start of the pandemic. The NA sequences from H1N1 (2009) were isolated in 23 different countries (Australia, Brazil, Canada, Chile, China, Denmark, Finland, France, Germany, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Philippines, Russia, Sweden, Thailand, United Kingdom, and the United States; the sequences from the United States represented 45 states. The NA sequences from H1N1 (2009) used from this study were isolated in 16 different countries (Brazil, Canada, China Columbia, France, Germany, Italy, Japan, Kazakhstan, Korea, Mexico, Nicaragua, Russia, Sweden, Thailand, and the United States; the sequences from the United States represented 36 different states. Pairs of epidemiologically matched HA and NA sequences (N = 92) were chosen to represent the same US state (or the same country in the case of non-US sequences) as close as possible to the same date (see www.biol.sc.edu/~austin). In 40 of 92 cases (43.5%), the paired sequences were from the same date; 65 of 92 (70.7%) were within 1 week of each other; and in no case were the two as much as 4 weeks apart. When more than one sequence was available for a given date and location, one sequence was selected at random for use in the matched pairs.

Statistical Methods

The sequences (Supplementary Tables 1 and 2, www.biol.sc.edu/~austin) were translated and aligned using the Prank program (*1*), and the alignment was imposed on the DNA sequences. Sequences containing undetermined nucleotides, premature stop codons and /or gaps were excluded from the analysis; partial sequences and laboratory strain sequences were also excluded from the analysis. Using the MEGA 4.0 program (*2*), we calculated the number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) by the method of Li (*3*). This method was used because it takes into account the effect of transitional bias, which is particularly important in the case of 2-fold degenerate sites (*3*). The synonymous nucleotide diversity (symbolized π_S) is defined as the mean of d_S for all pairwise comparisons among a set of sequences, while the nonsynonymous nucleotide diversity (symbolized π_N) is the mean of d_N for all pairwise comparisons among a set of sequences. Standard errors of π_S and π_N were estimated by the bootstrap method (*4*).

In each of the viruses, gene diversity ("heterozygosity") was estimated at each polymorphic site by the formula:

$$1 - \sum_{i=1}^{n} x_i^2$$

where n is the number of alleles and x_i is the frequency of the i^{th} allele in the set of sample sequences analyzed (*5,* p. 177). In coding regions, single-nucleotide polymorphisms (SNPs) were classified either as synonymous or nonsynonymous depending on their effect of the encoded amino acid sequence. Ambiguous sites were excluded from these analyses. The latter included sites at which both synonymous and nonsynonymous variants occurred in the set of sequences analyzed. Also excluded were certain polymorphic sites within codons with two or more polymorphic sites, when the polymorphism could be considered synonymous or nonsynonymous depending on the pathway taken by evolution. (For example, consider the two codons CTA and TTT. A mutation C→T in the first position would be synonymous if there were A in the third position, but not if there were T in the third position). Comparing gene diversities at synonymous and nonsynonymous polymorphic site provides evidence of ongoing purifying selection against slightly deleterious variants present in a population, since purifying selection will reduce the

frequency of slightly deleterious nonsynonymous variants in comparison to synonymous variants in the same genes (*6–9*).

Gene diversities were not normally distributed. Therefore, in testing for differences in mean gene diversity between synonymous and nonsynonymous SNP sites, randomization tests were used. In each test, 1,000 pseudo-datasets were created by sampling (with replacement) from the data; a difference between two categories was considered significant at the α level if it was greater than the absolute value of 100(1-α) % of the differences observed between the same categories in the pseudo-datasets.

The amino acid positions in HA at which a residue not seen in our sample of H1N1 (pre-2009) was fixed (i.e., at 100% frequency) in our sample of H1N1 (2009) were mapped on the crystal structure of HA from H1N1 (2009) (*10*) using RasTop version 2.2 (www.geneinfinity.org/rastop/).

References

1. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A. 2005;102:10557–62. PubMed DOI: 10.1073/pnas.0409137102

2. Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 2007;24:1596–9. PubMed DOI: 10.1093/molbev/msm092

3. Li W-H. Unbiased estimates of the rates of synonymous and nonsynonymous substitution. J Mol Evol. 1993;36:96–9. PubMed DOI: 10.1007/BF02407308

4. Nei M, Kumar S. Molecular evolution and phylogenetics. New York: Oxford University Press; 2000.

5. Nei M. Molecular evolutionary genetics. New York: Columbia University Press; 1987.

6. Hughes AL. Micro-scale signature of purifying selection in Marburg virus genomes. Gene. 2007;392:266–72. PubMed DOI: 10.1016/j.gene.2006.12.038

7. Hughes AL. Near neutrality: leading edge of the neutral theory of molecular evolution. Ann N Y Acad Sci. 2008;1133:162–79. PubMed DOI: 10.1196/annals.1438.001

8. Hughes AL. Small effective population sizes and rare nonsynonymous variants in potyviruses. Virology. 2009;393:127–34. PubMed DOI: 10.1016/j.virol.2009.07.016

9. Hughes AL, Packer B, Welsch R, Bergen AW, Chanock SJ, Yeager M. Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc Natl Acad Sci U S A. 2003;100:15754–7. PubMed DOI: 10.1073/pnas.2536718100

10. Xu R, Ekiert DC, Krause JC, Hai R, Crowe JE, Wilson IA. Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. Science. 2010; 328:357–60.

Table. Synonymous and nonsynonymous nucleotide diversity in 6 genes of pandemic (H1N1) 2009 virus*

Gene	No. sequences	π_S ± S.E.	π_N ± S.E.
PB1	325	0.0017 ± 0.0003	0.0004 ± 0.0001†
PB2	197	0.0902 ± 0.0070‡	0.0052 ± 0.0005‡
PA	171	0.0884 ± 0.0009‡	0.0044 ± 0.0004‡
NP	103	0.0946 ± 0.0080‡	0.0094 ± 0.0012‡
NS1	129	0.0957 ± 0.0131‡	0.0231 ± 0.0028‡
M1	50	0.2030 ± 0.0269‡	0.0171 ± 0.0037‡

*Tests of the hypothesis that π_S or π_N equals the corresponding value in hemagglutinin (contact A.L.H. for Supplementary Table 1). There was a significant difference between π_S and π_N values and the corresponding values for neruaminidase (contact A.L.H. for Supplementary Table 1; p<0.001) in every case. There was a significant difference (p<0.001) between π_S and π_N in every gene. π_S, *synonymous;* π_N, nonsynonymous; PB, polybasic protein; PA, polyacidic protein; NP, nucleocapsid protein; NS, nonstructural protein; M, matrix.
†p<0.05.
‡p<0.001.