# Genome Hotspots for Nucleotide Substitutions and the Evolution of Influenza A (H1N1) Human Strains

Alberto Civetta[1,*], David Cecil Murphy Ostapchuk[2], and Basil Nwali[1,3]

[1]Department of Biology, University of Winnipeg, Winnipeg, Manitoba R3B 2G3, Canada

[2]Department of Physics, University of Winnipeg, Winnipeg, Manitoba, Canada

[3]Department of Biochemistry, Ebonyi State University, Abakaliki, Nigeria

*Corresponding author: E-mail: a.civetta@uwinnipeg.ca.

**Accepted:** March 7, 2016

## Abstract

In recent years a number of studies have brought attention to the role of positive selection during the evolution of antigenic escape by influenza strains. Particularly, the identification of positively selected sites within antigenic domains of viral surface proteins has been used to suggest that the evolution of viral–host receptor binding specificity is driven by selection. Here we show that, following the 1918 outbreak, the antigenic sites of the hemagglutinin (HA) viral surface protein and the stalk region of neuraminidase became substitution hotspots. The hotspots show similar patterns of nucleotide substitution bias at synonymous and nonsynonymous sites. Such bias imposes directionality in amino acid replacements that can influence signals of selection at antigenic sites. Our results suggest that the high accumulation of substitutions within the antigenic sites of HA can explain not only cases of antigenic escape by antigenic drift but also lead to occasional episodes of viral extinction.

**Key words:** influenza, antigenic drift, selection, substitution hotspots, extinction.

## Introduction

The first recorded human pandemic caused by the influenza A (H1N1) strain, known as the Spanish Flu, started at the end of the First World War and spread worldwide infecting and causing death of a large percentage of the human population. The virus spread to swine, and then in 1957 it went extinct. A strain similar to the ones circulating up until the 1950s reappeared in 1977, and in 2009 a swine version of H1N1 moved to the human population causing an outbreak (Kendal et al. 1978; Nakajima et al. 1978; Taubenberger and Morens 2006; Garten et al. 2009; Smith et al. 2009; Christman et al. 2011).

The influenza A virus consists of a single-stranded RNA genome divided into eight segments. The eight segments code for polymerase, matrix, and structural proteins. Two envelope glycoproteins, hemagglutinin (HA) and neuraminidase (NA), are responsible for attaching the virus to the host and allowing the progression of infection (Szewczyk et al. 2014). HA is directly involved in the attachment of the virus to the host cell receptors and mediating fusion; and the antigenic sites of HA have been empirically identified using antibodies (Gerhard et al. 1981; Caton et al. 1982; Gamblin et al. 2004). The main function of NA is exerted through the enzymatic activity of the protein's globular domains, which cleaves sialic acid residues that otherwise keep the viral particles attached to the host cells (Air 2012; Marcelin et al. 2012; Wohlbold and Krammer 2014).

The origin and spread of particular genome variants facilitating viral escape from the host immune system have been explained by antigenic drift. Antigenic drift involves the accumulation of mutations within antibody-binding sites that render the virus to be less effectively targeted by the host immune system, thereby facilitating the spread of the infection. Antigenic drift posits that a constant rate of accumulation of substitutions (mutations) over time can result in antigenic variants that enable the virus to escape the immune system. Under antigenic drift, selection also plays a role because the strength of the host immune response and the severity and duration of the epidemic influences the rate of antigenic change (Kilbourne et al. 1990; de Jong et al. 2000; Sandbulte et al. 2011). In recent years, different studies have predicted sites within the influenza HA and NA proteins as likely influencing the ability of the virus to adapt to the host

© The Author 2016. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

immune response (Suzuki 2006; Shen et al. 2009; Bhatt et al. 2011; Li et al. 2011). In particular, the identification of selected (directional/positive) sites at antigenic positions suggests that adaptations to counteract the host response are a major force driving rates of amino acid substitutions.

We propose that hypotheses concerning the role of antigenic drift and positive selection in fuelling substitution rates at antigenic sites underestimate the effectiveness of purifying or negative selection in removing deleterious mutations. Moreover, the efficacy of either positive or negative selection is highly dependent on population size, as chance can drift deleterious mutations to fixation (Li and Graur 1991; Hartl and Clark 2007). The 1918 pandemic mortality happened in three waves from 1918 to 1919, after which the strain circulated in the human population with occasionally less severe epidemics. For example, during the 1930s and 1940s outbreaks, the percentage of deaths not due to influenza increased from 8% in 1918–1919 to 85% (Collins and Lehmann 1953; Taubenberger and Morens 2006). Such a reduction in the severity of the epidemic coupled with periods, during which the strain circulated in the human population, suggests a reduction in selective pressures relative to the original pandemic of 1918–1919. Weaker selective constraints as the epidemic is brought under control coupled with variation in mutational pressure across the viral genome could provide a better explanation for rapid evolution of the influenza virus.

Here, we effectively test for the occurrence of substitution hotspots within segments of the influenza genome. We examine sequences of the "Spanish" flu strains circulating from the first reported outbreak of 1918 until its effective extinction in 1957. The 1918–1957 period was purposely chosen for two reasons: 1) It expands the period subsequent to the Spanish flu outbreak up to the effective extinction of the strains, 2) It predates its 2000s re-emergence, which has led to claims on the role of positive selection as driving antigenic turnover. We find nucleotide substitution hotspots within both HA and NA. Interestingly the substitution hotspots for HA are located within the receptor binding subdomain and overlap with sites claimed to evolve under positive selection after the 2000s viral re-emergence (Shen et al. 2009). Nucleotide substitution biases within the hotspots affect both synonymous and nonsynonymous sites and impose directionality on amino acid changes. We argue that these hotspots evolve by differential mutation accumulation coupled with relaxation of selective constraint. Given the neutral to deleterious effect of most mutations, the accumulation of variants at these antigenic sites in strains circulating between 1918 and 1957 can more parsimoniously explain both the extinction episode as well as later cases (post 1950s) of antigenic escape among surviving viral strains.

## Results

We downloaded 743 complete sequences for all genome segments of the A (H1N1) influenza viral strains affecting humans during the 1918–1957 period. We found evidence of substitution hotspots at specific genome segment sites for the two surface glycoproteins HA and NA, while substitution coldspots were detected for HA and NP proteins (fig. 1). Overall, both the HA and NA segments that code for surface glycoproteins show nucleotide composition bias at the first codon position with high A and low C content (HA: $\chi^2 = 9.11$; df = 3; $P = 0.028$; NA: $\chi^2 = 12.94$; df = 3; $P = 0.005$). For HA, two hotspots were identified. The first hotspot is located from position 502–543 (13 codons) and shows a strong deviation from a random nucleotide substitution pattern ($\chi^2 = 30.00$; df = 5; $P < 0.001$) with A-G mutations making for 50% of all changes observed (fig. 2*A*). The proportion of A-G substitutions per synonymous sites (0.58) is similar to the proportion of such substitutions at nonsynonymous sites (0.42) (Fisher-exact test: $P = 0.489$), and the result is also nonsignificant if comparisons are made between first or second (0.50) versus third (0.43) codon positions (Fisher-exact test: $P = 0.748$) (supplementary table S1, Supplementary Material online). The other HA hotspot (598–624) shows a similar pattern of nonrandom nucleotide substitutions, although the deviation is weaker ($\chi^2 = 11.00$; df = 5; $P = 0.051$), with A-G and A-C mutations making up 33% and 24% of all substitutions respectively (fig. 2*B*). Among A-G substitutions, the proportions are similar at nonsynonymous versus synonymous (Fisher exact test: $P = 1.000$) or first–second versus third codon position (Fisher exact test: $P = 0.692$) (supplementary table S2, Supplementary Material online). One substitution hotspot in NA from sites 133 to 171 shows a larger proportion of A-G and A-C substitutions, although the deviation from random nucleotide substitution pattern is not statistically significant ($\chi^2 = 7.70$; df = 7; $P = 0.360$) (fig. 2*C*). The proportion of A-G changes at identifiable synonymous versus nonsynonymous sites are not significantly different (0.20 and 0.23, respectively; Fisher-exact test: $P = 1.000$). Similarly, no significant differences were found when the proportion of A-G substitutions were compared between first and second (0.27) versus the third (0.08) codon positions (Fisher-exact test: $P = 0.229$) (supplementary table S3, Supplementary Material online). In summary, the fact that substitution bias at hotspots is similar at both synonymous and nonsynonymous sites indicates that such substitutions are not driven by selective changes for amino acid turn over. In fact, an analysis of $d_N$ and $d_S$ estimates shows that during the 1918–1957 period both genes have overall evolved under negative or purifying selection, with no evidence of selection within the nucleotide substitution hotspots (table 1). Interestingly, this neutrally driven bias in substitutions might impose directionality in amino acid changes.

The two HA hotspots map within the cell receptor binding subdomain and span several previously identified Sa and Sb antigenic sites (fig. 3). Sites previously identified as under directional selection in human H1N1 strains circulating in 2006–2008, and particularly positively selected amino acids 156 and 190, are located within the identified hotspots (fig. 3). Substitution hotspots show nucleotide substitution bias that

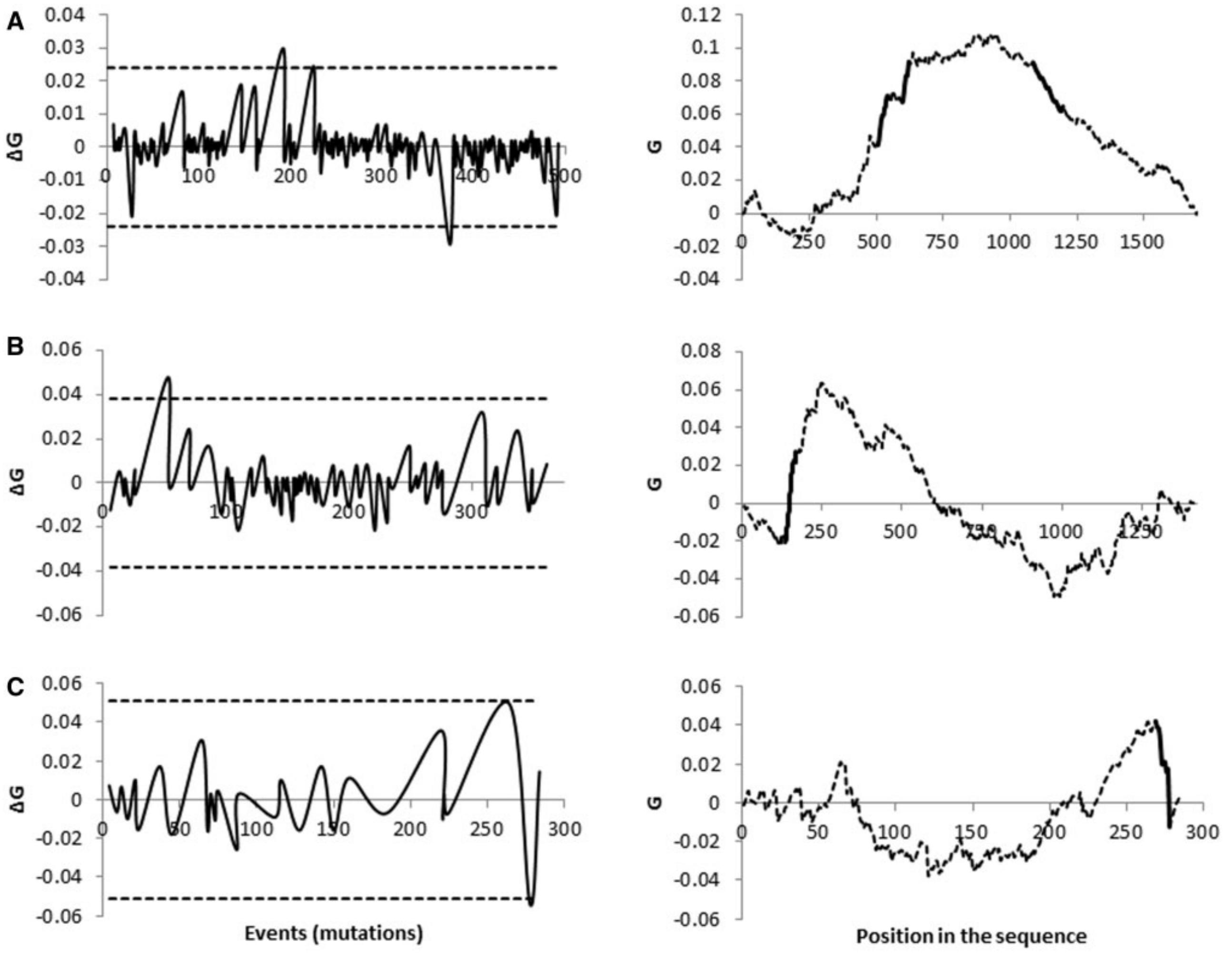


FIG. 1.—Plots of $G$ scores between nucleotide changes (events) and the differential accumulation of events ($\Delta G$) along gene sequences. The dashed lines indicate $G$ scores that identify both hot (positive $G$) and cold (negative $G$) substitution spots (statistical threshold: $P < 0.05$). The position of hot and cold spots in the gene segment sequence is also shown as solid lines in $\Delta G$ plots. Plots are shown for HA ($A$), NA ($B$) and NP ($C$) respectively.

similarly affects degenerate and nondegenerate codon sites (fig. 4 and supplementary tables S1–S3, Supplementary Material online). Such bias drives amino acid changes at NA and HA hotspots as shown by the predominance of substitutions involving amino acids changes driven by A-G replacements at the first and second codon position (fig. 5). These results, along with the fact that average estimates of $d_N$ and $d_S$ within nucleotide substitution hotspots do not significantly differ from each other (table 1), support mutation-driven evolution during the 1918–1957 period, leading to the origin of variants that can either drive extinction or confer antigenic drift ability.

## Discussion

We found nucleotide substitution hotspots within viral strains of H1N1 A circulating in the human population during the 1918–1957 period. The hotspots were localized within the two surface glycoproteins HA and NA. Such hotspots showed patterns of nucleotide substitution bias, with a significant over representation of A-G transitions. Substitution biases have been previously documented across a large range of vial genomes. A few examples of different biases are the hepatitis C virus genome with twice as many T-C than A-G transitions (Smith and Simmonds 1997), the maize streak virus (MSV) with an overrepresentation of G-T transversions (van der Walt et al. 2008) and the begomovirus tomato yellow leaf curl virus with high rates of C-T and G-A transitions (Duffy and Holmes 2008). There can often be different molecular basis for such biases in substitution rates. In the case of H1N1 A, there is evidence of an overtime increase of A and T content along with a decrease in G and C as the virus evolved in a human host. This trend was shown by the authors to be explained by a neutral process of host-dependent mutation bias (Rabadan et al. 2006). Using whole gene segment sequences, Rabadan et al. (2006) did not detect the same

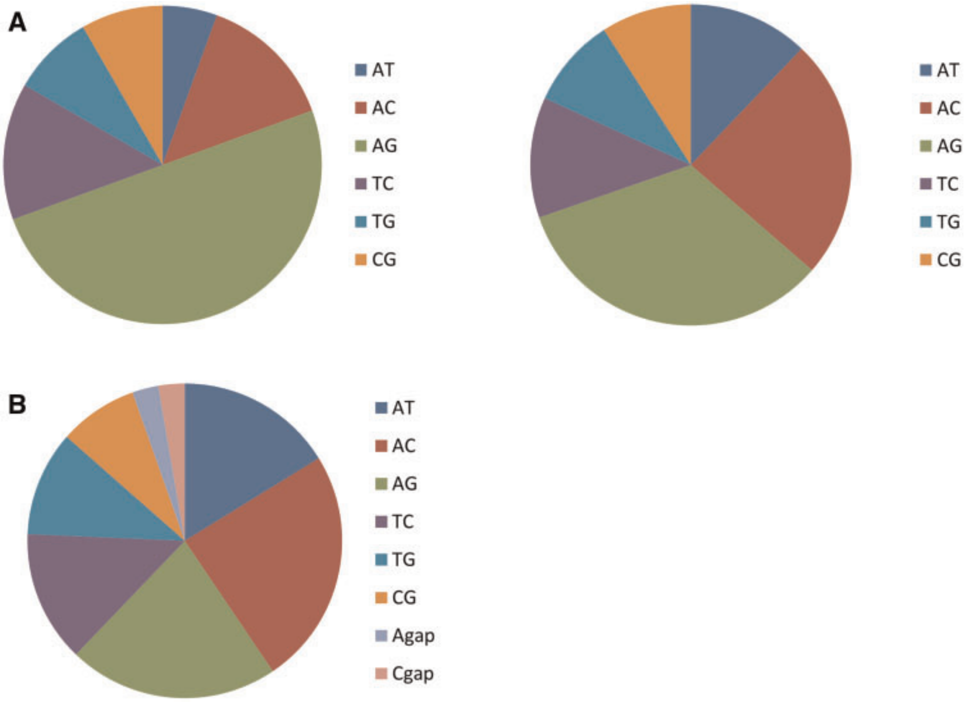


FIG. 2.—Substitution patterns within HA (*A*) and NA (*B*) hotspots.

**Table 1**
Test of Selection at HA and NA Genes as well as Nucleotide Substitution Hotspots within the Genes

| Gene | $d_N - d_S \pm$ SE | *P* value |
|---|---|---|
| HA (whole gene) | −0.067 ± 0.009 | <0.001 (negative selection) |
| HA (hotspot 1: 502–543) | 0.094 ± 0.113 | 0.375 (neutral) |
| HA (hotspot 2: 598–624) | 0.041 ± 0.128 | 0.731 (neutral) |
| HA (combined: 502–624) | 0.039 ± 0.040 | 0.333 (neutral) |
| NA (whole gene) | −0.103 ± 0.013 | <0.001 (negative selection) |
| NA (hotspot: 133–171) | 0.041 ± 0.051 | 0.409 (neutral) |

pattern of nucleotide biases for HA and NA, which they speculated to be a consequence of selection counteracting the effects imposed by mutation bias. Our results show localized biases in nucleotide substitutions within the HA and NA substitution hotspots. It is unclear what mechanism might cause the detected bias. Our results show that biases in substitution are overall neutral, so it is unlikely that selective pressures within the host cellular environment drive the observed pattern. The bias in substitutions is not different across codon positions, indicating that it is not driven by translational efficiency that could favor adenine content at the third nucleotide position of codons. Reverse transcription has been shown to most likely induce patterns of A-G bias in retroviruses (Müller and Bonhoeffer 2005), but reverse transcriptase is not involved in influenza viral replication. We speculate that two possible explanations for the observed pattern of substitutions are error bias induced by the viral RNA-polymerase during the copying of viral RNA in the cellular host environment or deamination by the human host APOBEC3G protein. Interestingly, APOBEC3G expression has been shown to be induced in human cells upon influenza infection without affecting the propagation of the virus (Pauli et al. 2009).

The NA protein has an N-terminal cytoplasmic domain, a transmembrane domain, a stalk of variable length, and a globular head domain that contains the enzymatic active site (Air 2012). The hotspot found in NA is located within the stalk domain, a region that has been previously identified as hypervariable, with deletions being common. A 20 amino acid deletion within the NA stalk region of wild duck influenza virus has been associated with viral transmission from waterfowl to chicken (Munier et al. 2010). In terms of infection progression, the main role of NA is to free virions from the protective mucin layer allowing access to the underlying respiratory epithelium (Wohlbold and Krammer 2014). The globular domain within NA mediates such processes and positively selected sites have been mapped within the globular domain of NA (Li et al. 2011). We found no evidence of hotspots within the globular head domain that confers enzymatic activity. While deletion of the stalk region might facilitate viral transmission, it appears that the stalk can simply accumulate neutral or slightly deleterious mutations without affecting the main function of the protein during infection.

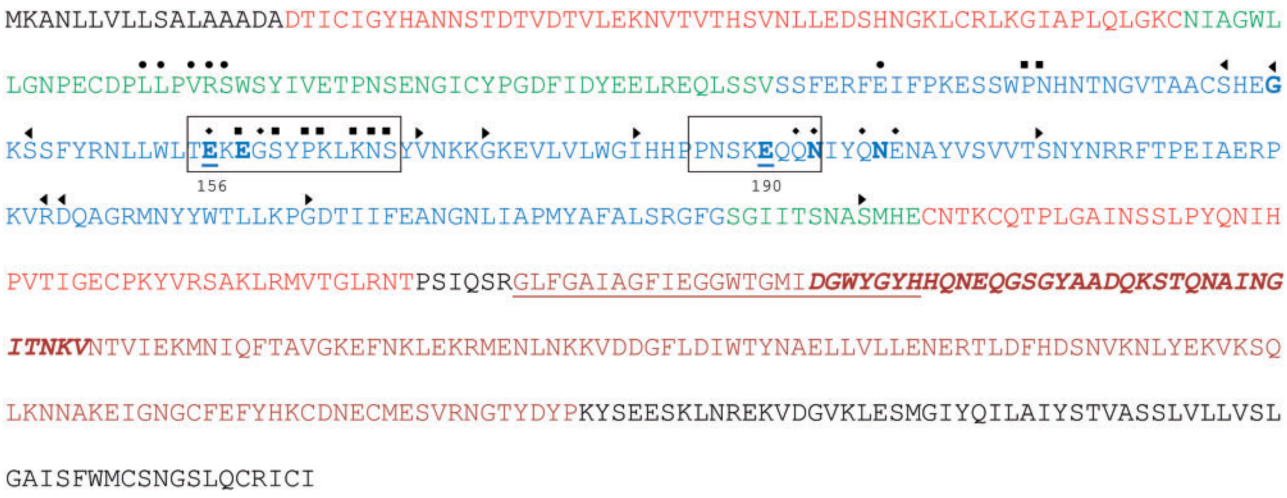

MKANLLVLLSALAAADADTICIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDSHNGKLCRLKGIAPLQLGKCNIAGWL
LGNPECDPLLPVRSWSYIVETPNSENGICYPGDFIDYEELREQLSSVSSFERFEIFPKESSWPNHNTNGVTAACSHEG
KSSFYRNLLWLTEKEGSYPKLKNSYVNKKGKEVLVLWGIHHPPNSKEQQNIYQNENAYVSVVTSNYNRRFTPEIAERP
156
190
KVRDQAGRMNYYWTLLKPGDTIIFEANGNLIAPMYAFALSRGFGSGIITSNASMHECNTKCQTPLGAINSSLPYQNIH
PVTIGECPKYVRSAKLRMVTGLRNTPSIQSRGLFGAIAGFIEGGWTGMIDGWYGYHHQNEQGSGYAADQKSTQNAING
ITNKVNTVIEKMNIQFTAVGKEFNKLEKRMENLNKKVDDGFLDIWTYNAELLVLLENERTLDFHDSNVKNLYEKVKSQ
LKNNAKEIGNGCFEFYHKCDNECMESVRNGTYDYPKYSEESKLNREKVDGVKLESMGIYQILAIYSTVASSLVLLVSL
GAISFWMCSNGSLQCRICI

**Fig. 3.**—Amino acid sequence of the HA protein of influenza virus (A/Puerto Rico/8/34/Mount Sinai. Accession number AAM75158). Colored amino acid residues indicate the subdomain structure of HA (red = membrane fusion subdomain F′ (HA1 peptide); green = vestigial esterase subdomain; blue = receptor binding subdomain; brown = membrane fusion subdomain F (HA2 peptide); brown underlined = HA2 fusion peptide). Antigenic sites are labeled as in Caton et al. (1982) (circle = Cb; square = Sa; rhomboid = Sb; triangles = Ca2 (left pointing) or Ca1 (right pointing)). Substitution hotspots are boxed. Amino acid sites previously identified by Shen et al. (2009) as evolving under directional selection (bold) or positive selection (underlined: 156 and 190) are identified. The coldspot is within the HA2 membrane fusion subdomain F (brown) and shown in bold italics.

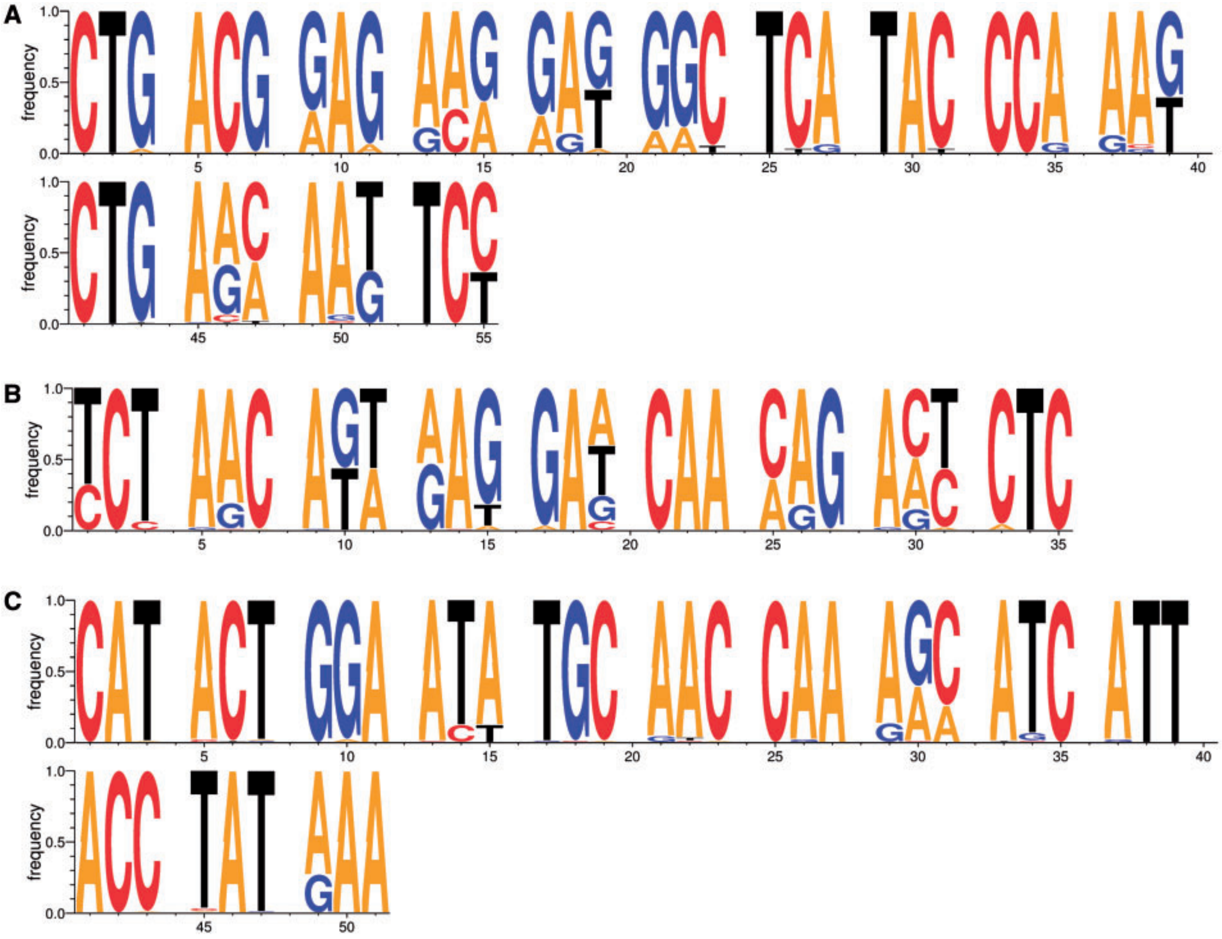
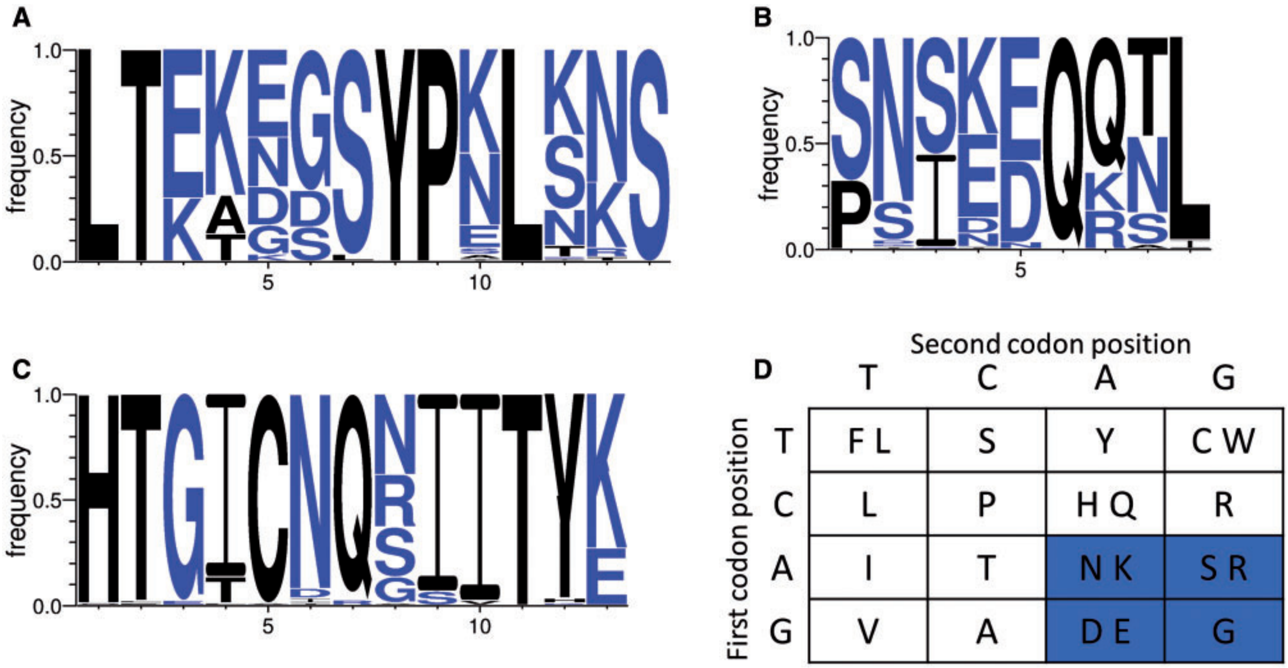


**Fig. 4.**—Nucleotide substitutions biases at substitution hotspots for HA 502 to 543 (*A*), HA 598 to 624 (*B*) and NA 133 to 171 (*C*).



| | T | C | A | G |
|---|---|---|---|---|
| T | F L | S | Y | C W |
| C | L | P | H Q | R |
| A | I | T | N K | S R |
| G | V | A | D E | G |

**Fig. 5.**—Amino acid substitution patterns for all three (*A–C*) hotspots as in figure 4. Blue is used for amino acids with A or G in the first and second codon position. The preponderance of amino acid substitutions driven by A-G nucleotide substitution biases at first and second codon positions (*D*) is highlighted in blue.

HA binds and fuses to host cells by recognition of cell surface receptors. Changes in binding specificity are crucial in allowing either cross-species transfer of infections or affecting the ability to bind and infect hosts (Gamblin et al 2004). The first crucial step of viral infection is binding to sialic acid cell receptors mediated by the HA receptor binding protein subdomain. After endocytosis of viral particles, cleavage of HA by host proteases into HA1 and HA2 peptides and extrusion of the HA2 fusion peptide become key events in mediating cell membrane fusion, which allows replication and production of new viruses. The fusion peptide, located at the amino end of HA2, is essential for membrane fusion (Mair et al. 2014). The HA coldspot identifies a highly conserved region that partially overlaps with the HA2 fusion peptide (fig. 3) and is likely preserved due to its functional relevance in mediating cellular membrane fusion.

The two HA hotspots that map within the HA protein receptor subdomain include several known antigenic sites. What is striking about the genome location of these hotspots is that the molecular bias in mutational changes predates previously identified episodes of positive selection at the same antigenic sites. The rapid accumulation of changes at antigenic sites from 1918 to 1957 could have caused a major decrease in the viral population due to the inability of the virus to purge slightly deleterious changes accumulating in their genomes (Muller 1963; Gabriel and Bürger 1993; Lynch et al. 1993). While the accumulation of deleterious mutations could eventually result in extinction of the virus through what it has been referred to as natural attenuation (Carter and Sanford 2012), relaxed negative selection pressures due to decreases in viral population size would have allowed for drift and an increase in frequency of variants that could later become of adaptive value. Our results suggest that the outcome of evolution of influenza strains is highly dependent on the nonrandom distribution of substitutions with a particular overrepresentation of biased nucleotide substitution within protein regions with antigenic roles.

## Materials and Methods

Complete genome segment nucleotide sequence data was downloaded from the Influenza Research Database (www.fludb.org) for A (H1N1) influenza strains affecting humans from 1918 to 1957. The entire coding regions of all eight genome segments were aligned using MUSCLE within MEGA (Tamura et al. 2013). Sequence alignments are available upon request. We transformed the alignments using MACML (Zhang and Townsend 2009) into a string of 0s and 1s, were 1s denote the occurrence of substitutions while 0s identify monomorphic sites. We used a PERL script to count the number of events (1s) in any given alignment as well as the position at which such events occurred.

We used the test proposed by Tang and Lewontin (1999) to detect differential variability in the accumulation of substitutions across the genome segments (i.e. substitution hotspots or coldspots). Briefly, the method tests for significant deviations from a uniform distribution of changes (events) using an empirical cumulative distribution function. Given a sequence of length $N$ with $n$ changes at positions $x_k$ (with $k$ ranging from 1 to $n$), deviations from a uniform deviation of events is measured by the $G$ function. $G$ is calculated as the difference between the relative occurrences of the nucleotide changes ($k/n$) minus their relative position in the sequence ($x_k/N$). Differences between the values of the $G$ function at two events ($\Delta G$) measures the differential accumulation of nucleotide substitutions. To test whether any region within the sequence deviates from a random accumulation of changes, a $T$-statistic is defined as the $\Delta G$ with the highest absolute value. A uniform distribution of changes is rejected if $T$ is higher than $T^*$. $T^*$ is a null random distribution of $T$ obtained using Monte Carlo simulations to produce 100,000 samples of $n$ events by sampling sites without replacement along a sequence of length $N$. We develop source code in C (available upon request) for the implementation of the Tang–Lewontin method.

Once hotspots and coldspots were detected, we analyzed nucleotide composition and substitution patterns using both DNASp v5.10 (Librado and Rozas 2009) and MEGA v6.0 (Tamura et al. 2013). Plots of relative frequency of amino acids or nucleotides within substitution hotspots were done using WebLogo 3 (http://weblogo.threeplusone.com/create.cgi, last accessed March 21, 2016) (Crooks et al. 2004). Test of selection were conducted using $Z$-test for differences between $d_N$ and $d_S$ estimates (test of neutrality). We used the modified Nei–Gojobori (Jukes Cantor) method within MEGA for estimates of $d_N$ and $d_S$, as it corrects to account for biases in substitution rates and for multiple substitutions at the same site.

## Supplementary Material

Supplementary tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

## Acknowledgments

We would like to thank Dr Renee Douville and Dr Sara Good for their comments on the paper. Dr Jeff Martin was helpful during the development of code for the implementation of the Tang–Lewontin method for detection of substitution cold and hot spots. B.N. was supported by a visiting doctoral scholarship from Ebonyi State University. This work was funded by an individual Discovery grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada to A.C.

## Literature Cited

Air GM. 2012. Influenza neuraminidase. Influenza Other Respir Viruses 6:245–256.

Bhatt S, Holmes EC, Pybus OG. 2011. The genomic rate of molecular adaptation of the human influenza A virus. Mol Biol Evol. 28:2443–2451.

Carter RW, Sanford JC. 2012. A new look at an old virus: patterns of mutation accumulation in the human H1N1 influenza virus since 1918. Theor Biol Med Model. 9:42.

Caton AJ, Brownlee GG, Yewdell JW, Gerhard W. 1982. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). Cell 31:417–427.

Christman MC, Kedwaii A, Xu J, Donis RO, Lu G. 2011. Pandemic (H1N1) 2009 virus revisited: an evolutionary retrospective. Infect Genet Evol. 11:803–811.

Collins SD, Lehmann J. 1953. Excess deaths from influenza and pneumonia and from important chronic diseases during epidemic periods, 1918-51. Public Health Monogr. 10:1–21

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res. 14:1188–1190.

de Jong JC, Rimmelzwaan GF, Fouchier RAM, Osterhaus ADME. 2000. Influenza virus: a master of metamorphosis. J Infect. 40:218–228.

Duffy S, Holmes EC. 2008. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. J Gen Virol. 90:1539–1547.

Gabriel WLM, Bürger R. 1993. Muller's ratchet and mutational meltdowns. Evol Int J Org Evol. 47:1744–1757.

Garten RJ, et al. 2009. Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science 325:197–201.

Gamblin SJ, et al. 2004. The structure and receptor binding properties of the 1918 influenza hemagglutinin. Science 303:1838–1842.

Gerhard W, Yewdell J, Frankel ME, Webster R. 1981. Antigenic structure of influenza virus haemagglutinin defined by hybridoma antibodies. Nature 290:713–717.

Hartl DL, Clark AG. 2007. Principles of population genetics. 4th ed. Sunderland, MA: Sinauer Associates.

Kendal AP, Noble GR, Skehel JJ, Dowdle WR. 1978. Antigenic similarity of influenza A (H1N1) viruses from epidemics in 1977–1978 to "Scandinavian" strains isolated in epidemics of 1950–1951. Virology 89:632–636.

Kilbourne ED, Johansson BE, Grajower B. 1990. Independent and disparate evolution in nature of influenza A virus hemagglutinin and neuraminidase glycoproteins. Proc Natl Acad Sci U S A. 87:786–790.

Li W-H, Graur D. 1991. Fundamentals of molecular evolution. Sunderland, MA: Sinauer Associates.

Li W, et al. 2011. Positive selection on hemagglutinin and neuraminidase genes of H1N1 influenza viruses. Virol J. 8:183.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452.

Lynch M, Burger R, Butcher D, Gabriel W. 1993. The mutational meltdown in asexual populations. J Hered. 84:339–344.

Mair CM, Ludwig K, Herrmann A, Sieben C. 2014. Receptor binding and pH stability - how influenza A virus hemagglutinin affects host-specific virus infection. Biochim Biophys Acta. 1838:1153–1168.

Marcelin G, Sandbulte MR, Webby RJ. 2012. Contribution of antibody production against neuraminidase to the protection afforded by influenza vaccines. Rev Med Virol. 22:267–279.

Muller HJ. 1963. The need for recombination to prevent genetic deterioration. Genetics 48:903–903.

Müller V, Bonhoeffer S. 2005. Guanine-adenine bias: a general property of retroid viruses that is unrelated to host-induced hypermutation. Trends Genet. 21:264–268.

Munier S, et al. 2010. A genetically engineered waterfowl influenza virus with a deletion in the stalk of the neuraminidase has increased virulence for chickens. J Virol. 84:940–952.

Nakajima K, Desselberger U, Palese P. 1978. Recent human influenza A (H1N1) viruses are closely related genetically to strains isolated in 1950. Nature 274:334–339.

Pauli EK, et al. 2009. High level expression of the anti-retroviral protein APOBEC3G is induced by influenza A virus but does not confer antiviral activity. Retrovirology 6:38.

Rabadan R, Levine AJ, Robins H. 2006. Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes. J Virol. 80:11887–11891.

Sandbulte MR, et al. 2011. Discordant antigenic drift of neuraminidase and hemagglutinin in H1N1 and H3N2 influenza viruses. Proc Natl Acad Sci U S A. 108:20748–20753.

Shen J, Ma J, Wang Q. 2009. Evolutionary trends of A(H1N1) influenza virus hemagglutinin since 1918. PLoS One 4:e7789.

Smith DB, Simmonds P. 1997. Characteristics of nucleotide substitution in the hepatitis C virus genome: constraints on sequence change in coding regions at both ends of the genome. J Mol Evol. 45:238–246.

Smith GJ, et al. 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459:1122–1125.

Suzuki Y. 2006. Natural selection on the influenza virus genome. Mol Biol Evol. 23:1902–1911.

Szewczyk B, Bieńkowska-Szewczyk K, Król E. 2014. Introduction to molecular biology of influenza A viruses. Acta Biochim Pol. 61:397–401.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 30:2725–2729.

Tang H, Lewontin RC. 1999. Locating regions of differential variability in DNA and protein sequences. Genetics 153:485–495.

Taubenberger JK, Morens DM. 2006. 1918 Influenza: the mother of all pandemics. Emerg Infect Dis. 12:15–22.

van der Walt E, Martin DP, Varsani A, Polston JE, Rybicki EP. 2008. Experimental observations of rapid Maize streak virus evolution reveal a strand-specific nucleotide substitution bias. Virol J. 5:104.

Wohlbold TJ, Krammer F. 2014. In the shadow of hemagglutinin: a growing interest in influenza viral neuraminidase and its role as a vaccine antigen. Viruses 6:2465–2494.

Zhang Z, Townsend JP. 2009. Maximum-likelihood model averaging to profile clustering of site types across discrete linear sequences. PLoS Comput Biol. 5:e1000421.

**Associate editor:** Chantal Abergel