

Research Article

Natural Selection Determines Synonymous Codon Usage Patterns of Neuraminidase (NA) Gene of the Different Subtypes of Influenza A Virus in Canada

Youhua Chen^{1,2}

¹ Department of Zoology, University of British Columbia, Vancouver, Canada V6T 1Z4

² Department of Renewable Resources, University of Alberta, Edmonton, Canada T6G 2H1

Correspondence should be addressed to Youhua Chen; haydi@126.com

Received 15 April 2013; Revised 1 July 2013; Accepted 25 August 2013; Published 2 January 2014

Academic Editor: Sílvia Bofill-Mas

Copyright © 2014 Youhua Chen. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Synonymous codon usage patterns of neuraminidase (NA) gene of 64 subtypes (one is a mixed subtype) of influenza A virus found in Canada were analyzed. In total, 1422 NA sequences were analyzed. Among the subtypes, H1N1 is the prevailing one with 516 NCBI accession records, followed by H3N2, H3N8, and H4N6. The year of 2009 has the highest report records for the NA sequences in Canada, corresponding to the 2009 pandemic event. Correspondence analysis on the RSCU values of the four major subtypes showed that they had distinct clustering patterns in the two-dimensional scatter plot, indicating that different subtypes of IAV utilized different preferential codons. This subtype clustering pattern implied the important influence of natural selection, which could be further evidenced by an extremely flattened regression line in the neutrality plot (GC12 versus G3s plot) and a significant phylogenetic signal on the distribution of different subtypes in the clades of the phylogenetic tree (λ statistic). In conclusion, different subtypes of IAV showed an evolutionary differentiation on choosing different optimal codons. Natural selection played a deterministic role to structure IAV codon usage patterns in Canada.

1. Introduction

Codon usage is not a random event [1]. Codon usage bias has been broadly observed, and different mechanisms have been proposed to explain the bias patterns, for example, mutation pressure, translational efficiency, gene length [2], dinucleotide bias [3], tRNA abundance [4], organ specificity [5], and so on. Codon usage bias patterns have been broadly studied in recent years, especially for virus genomes [3, 6, 7].

In recent years, codon usage patterns have been widely explored for influenza viruses [6, 8–12]. Among the three influenza viruses, influenza A virus (IAV) is the major concern since it has a lot of subtypes. IAV is a genus of Orthomyxoviridae family of viruses, which caused influenza in birds and mammals [6, 13]. Among the eight RNA segments of IAV, hemagglutinin and neuraminidase (NA) genes are the principal concerns.

Currently, most of modeling efforts on IAV are focused on Asian regions [6, 8]; little attention is paid on the evolutionary

patterns of IAV in local areas of North America. To fill such a knowledge gap, in the present study, I analyzed all the available 1436 NA ORFs for IAV found in Canada to reveal the codon usage patterns of IAV different subtypes in Canada.

2. Materials and Methods

2.1. Sequence Data. 1436 NA sequences found in IAV strains of Canada were extracted from NCBI GenBank database (<http://www.ncbi.nlm.nih.gov/>). Only the open reading frames (ORFs) are considered for codon usage bias analysis. The alignment of the large number of NA sequences is done using MAFFT software [14, 15]. The dataset and the alignment will be available at the online digital repository website: <http://datadryad.org/>.

2.2. Measure of Codon Usage Patterns. Relative synonymous codon usage values of each codon in a gene are calculated

to investigate the characteristics of synonymous codon usage. The RSCU index is calculated as follows [16]:

$$\text{RSCU} = \frac{g_{ij} \times n_j}{\sum_i^{n_j} g_{ij}}, \quad (1)$$

where g_{ij} is the observed number of the i th codon for the j th amino acid which has n_j kinds of synonymous codons. Higher (or lower) RSCU values, higher (or lower) frequencies of the codons being chosen. When the corresponding RSCU values of a codon are close to 1, it is used randomly and evenly.

2.3. Effective Number of Codons. The effective number of codons (ENCs) is a measure of bias from equal codon usage in a gene [17]. The calculation formula is

$$\text{ENC} = 2 + \frac{9}{\bar{F}_2} + \frac{1}{\bar{F}_3} + \frac{5}{\bar{F}_4} + \frac{3}{\bar{F}_6}, \quad (2)$$

where \bar{F}_k ($k = 2, 3, 4, 6$) is the mean of F_k values for the k -fold degenerate amino acids, which is estimated using the formula as follows:

$$F_k = \frac{nS - 1}{n - 1}, \quad (3)$$

where n is the total number of occurrences of the codons for that amino acid and

$$S = \sum_{i=1}^k \binom{n_i}{n}^2, \quad (4)$$

where n_i is the total number of occurrences of the i th codon for that amino acid.

ENC ranges from 20 for the strongest bias (where only one codon is used for each amino acid) to 61 for no bias (where all synonymous codons are used equally).

For revealing the relationship between GC3s and ENC values, the expected ENC values for different GC3s were calculated as follows:

$$\text{ENC}^{\text{expected}} = 2 + s + \frac{29}{s^2 + (1 - s)^2}, \quad (5)$$

where s denotes the value of GC3s [6]. These expected ENC values can form a unimodal curve. As such, the observed and expected ENC values can be compared to determine the influence of nucleotide compositional constraint on structuring synonymous codon usage bias. If the observed data points are on or nearby the null expected ENC curve, then compositional constraint plays the major role in structuring codon usage patterns. If the observed points are fallen below the null curve, then the role of natural selection emerges.

2.4. Correspondence Analysis and Correlation Analysis. Correspondence analysis (COA) was conducted to investigate the major trend involved in the codon usage patterns of the genomes from different virus strains, which were measured by corresponding RSCU values for the 59 codons (after excluding Met, Trp, and three termination codons) [7]. COA

is performed using “ca” package [18] under R environment [19].

The first two axes of COA (CA1 and CA2) were then subjected to correlation analysis. A Spearman’s rank correlation analysis was used to infer the relationships between different codon usage indices and the first two axes (CA1 and CA2) of IAV RSCU values.

2.5. Phylogenetic Signals of Different Subtypes of IAV for NA Gene. Phylogenetic signals revealed the clustering and overdispersion (or randomness) of the different subtypes in the phylogenetic tree. If different subtypes tend to group in the same clade, then the clustering pattern could be found and the phylogenetic signal should be detected. In contrast, if different subtypes tend to disperse across different clades, then the overdispersion pattern could be found and the phylogenetic signal should be minor. I implemented two metrics for testing phylogenetic signals for the major subtypes of IAV found in Canada: Blomberg’s K statistic [20] and Pagel’s λ statistic [21, 22]. Both tests were implemented in R [19] package “phytools” [23].

For inferring phylogenetic signals, a phylogenetic tree is required. Given that we have over 1400 NA sequences, we used the fastest software, FastTree [24], to construct a maximum likelihood (ML) tree, which is based on the sequences derived from the major subtypes (H1N1, H3N2, H3N8, and H4N6). The reconstructed ML tree will be available from the online digital repository website: <http://datadryad.org/>. The multichotomies of the tree are resolved using R [19] package “ape” [25], and the dating of divergence time is done using the mean path length method [26]. Negative branch lengths have occurred after molecular dating but do influence the inference of phylogenetic signal test.

3. Results and Discussion

3.1. Count and Temporal Distribution of Subtypes of Influenza A Viruses in Canada. As shown in Figure 1(a), the most prevailing subtype of IAV in Canada is H1N1 (with 516 records), followed by H3N2 (with 132 records), H3N8 (with 152 records), and H4N6 (with 113 records).

Based on strain records (Figure 1(b)), it was found that the year of 2009 has the highest number of IAV NA genes being sequenced with a number of 531, followed by the year of 2007 (with a number of 204) and the year of 1968 (66 records).

3.2. ENC-GC3s and Neutrality Plots. As shown in Figure 2, most points lay below the null curve considerably, while some points from H1N1 were very far from the null curve, indicating the importance of selection.

Neutrality plot of Figure 3(a) showed that natural selection should play a very crucial role in structuring the overall codon usage patterns across different subtypes of IAV in Canada. The relative influence of mutation pressure is very small on the first and second codon positions (GC12) with respect to that on the third codon position (GC3), as indicated by the slope of the regression line (0.0158). Such a conclusion is enforced since this regression slope is not

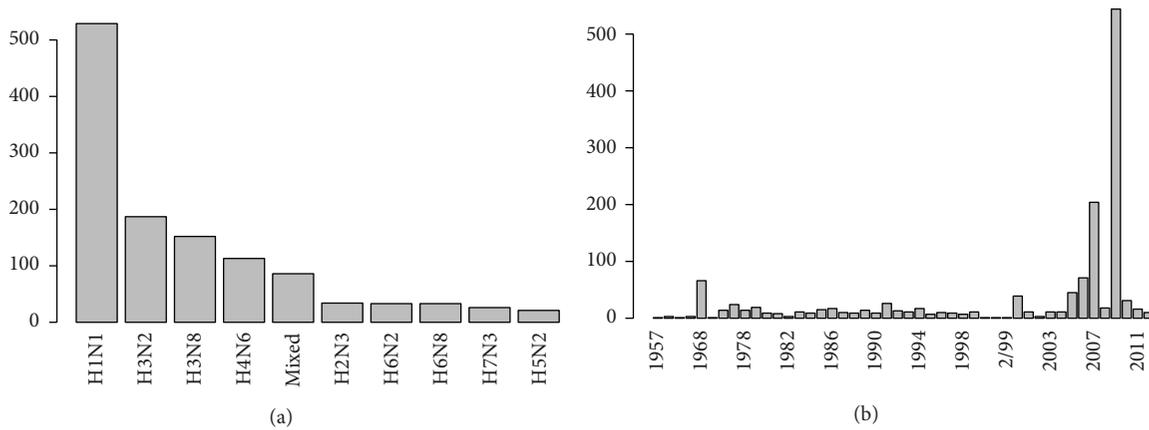


FIGURE 1: (a) Summary of the number of NA gene sequences for different subtypes of IAV found in Canada up to the year of 2013. (b) Summary of the number of NA gene sequences for different subtypes of IAV sequenced at each year in Canada up to the year of 2013.

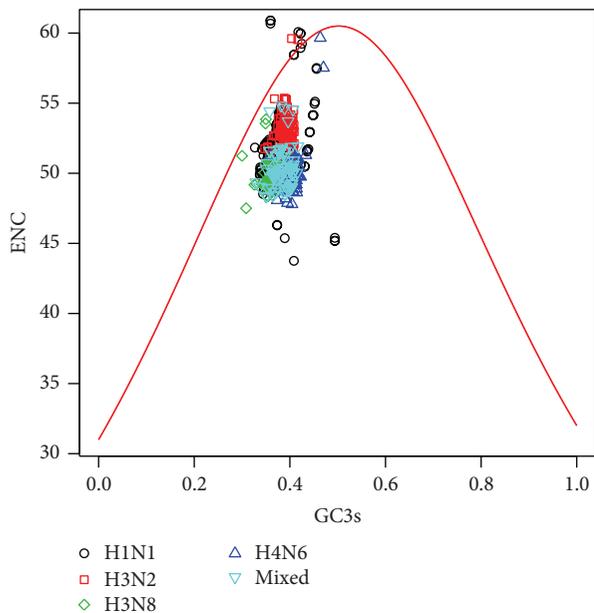


FIGURE 2: ENC-GC3s plot. The red curve indicated the null line of codon usage patterns, in which only mutational pressure is assumed to operate.

significant ($P > 0.05$), which suggests that there is no clear trend between GC12 and GC3 across different subtypes. Such a decoupling is argued to be structured by natural selection [27, 28].

Furthermore, when the neutrality plots are applied to each of the major subtypes (Figures 3(b), 3(c), 3(e), and 3(f)), most of them are found to have near-zero or negative regression slopes (indicated by correlation coefficients and associated P values). As such, natural selection is exclusively important to structure codon usage pattern of NA gene within a single subtype and between different subtypes of IAV in Canada.

3.3. Clustering Patterns of Codon Usage Patterns of Major Subtypes in Canada. As shown in the COA plot of the first two axes in Figure 4, apparently there are distinct clustering patterns among the first four subtypes of IAV. However, the mixed subtype group showed a dispersing pattern, which could be inferred that these mixed sequences are principally generated from subtypes H3N8 and H4N6 because most mixed sequences were well overlapped with the positions of these two known subtype groups.

With respect to the codon usage bias patterns of the four major subtypes (H1N1, H3N2, H3N8, and H4N6), it is found that they did show distinct preferences on choosing optimal codons (Table 1). For subtype H1N1, the most preferred codons (standard: RSCU > 1.2 being the highest in that subtype) were GCU, CUA, UUA, AGA, and AGU in comparison to other subtypes. For subtype H3N2, the most preferred codons were CAU, CCU, AGC, and UAU, respectively. For subtype H3N8, the most preferred codons were GCA, AAU, AAA, AAU, AGG, and ACU. For subtype H4N6, the most important codons were GAA, UUA, GGA, AUA, UUG, CCA, UCA, GUA, and GUG. Apparently, different subtypes tended to congruently use A-end codons in most cases. Such an observation is consistent with many previous studies working on specific subtypes of IAV [6, 8]. Interestingly, no most preferred codons were found for mixed subtype group (Table 1). This is completely consistent to the clustering patterns in Figure 4.

3.4. The Determinants of Codon Usage Patterns of Different Subtypes of IAV in Canada. As shown in the correlation between different codon usage indices and the first two axes of COA (Table 2), it is evidenced that the C3s and G3s individually are the principal drivers for determining the first major trend CA1 of codon usage bias patterns of IAV in Canada. Interestingly, GC3s, the combined G3s and C3s, is not a strong determinant of the first major trend CA1. Comparatively, A3s and A contents are the major predictors

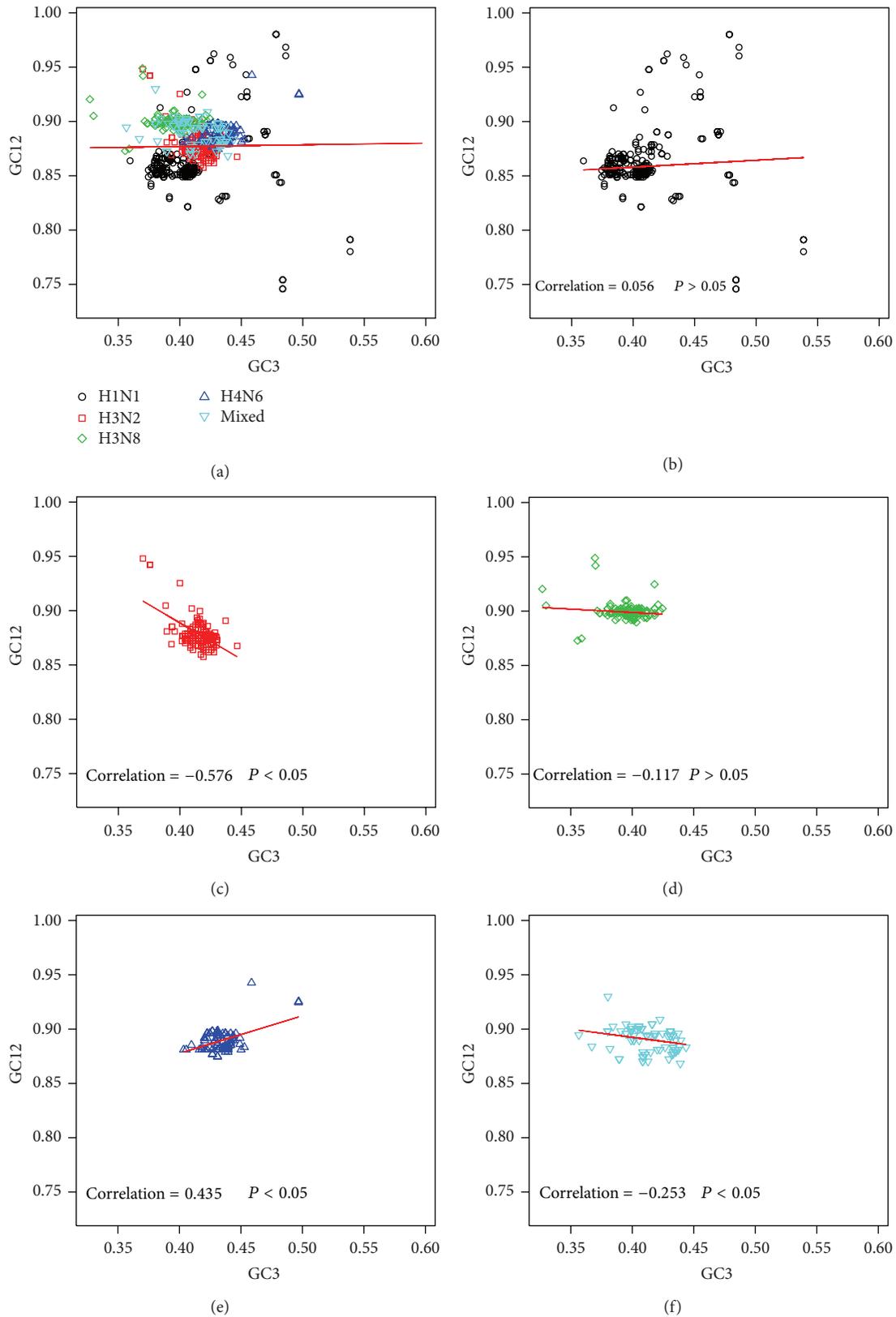


FIGURE 3: Neutrality plot between GC3 and GC12 for the NA gene of all and each of the major subtypes in Canada. (a) Neutrality plot for all major subtypes. The fitted regression line has the equation $GC12 = 0.0158 \times GC3 + 0.87$ ($R^2 = 0.0002, P = 0.592$). (b) Neutrality plot for subtype H1N1. (c) Neutrality plot for subtype H3N2. (d) Neutrality plot for subtype H3N8. (e) Neutrality plot for subtype H4N6. (f) Neutrality plot for mixed subtype.

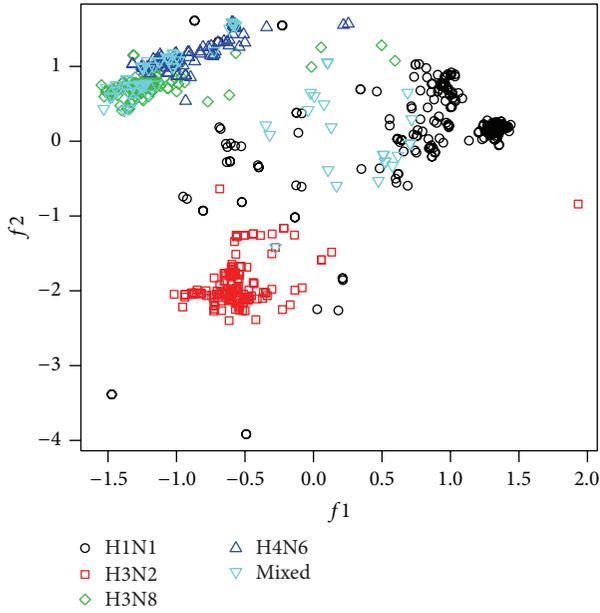


FIGURE 4: CA plot of RSCU values of NA sequences for different subtypes of IAV in Canada. Different colors indicated different subtypes. Clearly, sequences from the same subtypes tended to group in the same clusters, except for those for the mixed group.

of the second major trend of codon usage patterns of different subtypes of IAV.

3.5. Phylogenetic Clustering Patterns of Different Subtypes of IAV in Canada. Blomberg's K statistic failed to return a result because the phylogenetic covariance matrix based on the inferred ML tree is singular when the inverse of the matrix is taken. However, Pagel's λ statistic could be implemented. The randomization test (1000 runs) on the λ statistic for testing phylogenetic signals showed a significant result ($\lambda = 0.886$, log-likelihood = -1480.885 , and $P = 0$). The significant phylogenetic structure implied that NA sequences from the same subtype would tend to group in the same clade in the phylogenetic tree.

3.6. The Debate of Mutation versus Selection on IAV. Natural selection played a dominant role in structuring different subtypes of NA genes for IAV in Canada. The significant clustering patterns supported by correspondence analysis on the RSCU values and the phylogenetic signaling randomization test clearly support the evolutionary differentiation of NA genes of different subtypes of IAV in Canada. Neutrality plot further evidenced the dominance of natural selection on NA gene in Canada.

Our study, therefore, is different from previous studies on codon usage patterns of IAV because the previous studies showed that mutation selection should be prevailing [6, 8, 9]. Therefore, it is imperative to compare different subtypes found in different regions (either regional or continental levels) to better quantify the influence of environmental

TABLE 1: Codon usage comparison between the four major subtypes found in Canada using RSCU values. Only the codons with some RSCU values >1.2 across the four subtypes are considered for comparison of codon usage preference. The largest RSCU value for each of these codons is marked in boldface to see the most preferred subtype.

Codons	H1N1	H3N2	H3N8	H4N6	Mixed
GCA	1.141	1.483	2.487	2.142	2.116
GCC	0.777	0.857	0.719	0.512	0.665
GCG	0.233	0.394	0.22	0.552	0.328
GCU	1.848	1.266	0.574	0.793	0.891
UGC	0.979	0.998	1.053	1.175	1.114
UGU	1.021	1.002	0.947	0.825	0.886
GAC	0.96	0.895	0.785	0.897	0.861
GAU	1.04	1.105	1.215	1.103	1.139
GAA	1.27	1.017	1.246	1.306	1.203
GAG	0.73	0.983	0.754	0.694	0.797
UUC	1.003	0.952	0.941	1.203	1.048
UUU	0.997	1.048	1.059	0.797	0.952
GGA	1.563	1.317	2.016	2.118	1.986
GGC	0.605	0.763	0.537	0.422	0.459
GGG	0.981	0.937	0.811	1.082	0.954
GGU	0.85	0.983	0.636	0.378	0.601
CAC	0.731	0.255	0.522	0.378	0.516
CAU	1.269	1.745	1.478	1.622	1.484
AUA	1.365	1.215	1.088	1.423	1.311
AUC	0.643	0.764	0.501	0.896	0.668
AUU	0.992	1.021	1.411	0.681	1.021
AAA	1.029	1.313	1.259	1.256	1.179
AAG	0.971	0.687	0.741	0.744	0.821
CUA	1.406	0.962	1.269	1.283	1.309
CUC	0.349	0.882	0.88	0.64	0.705
CUG	0.794	0.833	0.633	0.635	0.779
CUU	0.444	1.089	0.828	0.675	0.642
UUA	1.419	0.616	0.968	0.938	0.943
UUG	1.587	1.617	1.422	1.83	1.621
AUG	1	1	1	1	1
AAC	0.844	0.748	0.534	0.867	0.683
AAU	1.156	1.252	1.466	1.133	1.317
CCA	1.91	0.932	1.815	1.951	1.829
CCC	0.602	0.703	1.165	0.441	0.791
CCG	0.402	0.267	0.301	0.672	0.42
CCU	1.086	2.097	0.719	0.935	0.959
CAA	1.101	1.19	1.07	0.897	1.027
CAG	0.899	0.81	0.93	1.103	0.973
AGA	3.389	2.763	3.071	2.95	3.023
AGG	1.238	2.084	2.616	2.108	2.222
CGA	0.924	0.37	0.032	0.472	0.285
CGC	0.299	0.257	0.003	0.009	0.068
CGG	0.056	0.502	0.277	0.461	0.379
CGU	0.095	0.023	0.002	0	0.023
AGC	1.133	1.209	0.806	1.164	1.034

TABLE 1: Continued.

Codons	H1N1	H3N2	H3N8	H4N6	Mixed
AGU	1.342	0.887	1.319	1.07	1.244
UCA	1.414	1.703	1.853	1.923	1.798
UCC	0.758	1.103	0.956	1.066	0.944
UCG	0.334	0.234	0.304	0.44	0.339
UCU	1.019	0.863	0.761	0.337	0.642
ACA	1.579	1.477	1.553	1.973	1.709
ACC	1.044	1.128	0.671	1.176	0.931
ACG	0.063	0.246	0.538	0.321	0.416
ACU	1.314	1.149	1.238	0.53	0.943
GUA	1.067	1.042	1.281	1.829	1.481
GUC	0.805	0.739	0.562	0.62	0.613
GUG	1.123	1.168	1.091	1.248	1.116
GUU	1.005	1.051	1.066	0.303	0.79
UGG	1	1	1	1	1
UAC	0.916	0.351	1.125	1.039	1.103
UAU	1.084	1.649	0.875	0.961	0.897

TABLE 2: The correlation between different codon usage indices and the first two axes of COA analysis. The most correlated variables for each axis of COA are marked in boldface. Significant correlations at the level of $P < 0.05$ are marked with asterisks.

	CA1	CA2
A	0.191007*	-0.63365*
C	-0.02447	0.10123*
G	0.523703*	-0.05803*
T	-0.47806*	0.482319*
GC	0.459962*	0.041116
A3s	0.37797*	-0.78633*
C3s	-0.55916*	0.21002*
G3s	0.689769*	-0.03151
T3s	-0.41266*	0.586561*
GC3s	0.144409*	0.161182*
ENC	-0.47381*	0.511762*

selection and mutational pressure on the evolution of virus genes.

4. Conclusions

Because (1) different subtypes of IAV have utilized different optimal codons, (2) flattened regression lines are always found in the neutrality plots, and (3) phylogenetic signal is significant for the distribution of subtypes over the phylogenetic tree. It is concluded that different subtypes of IAV in Canada show evolutionary differentiation patterns, and natural selection plays a central role in structuring codon usage patterns for IAV in the region.

Conflict of Interests

There is no conflict of interests regarding the materials of the study.

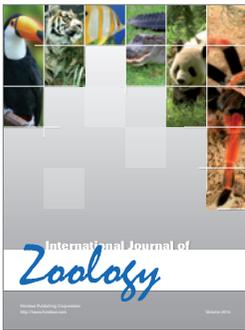
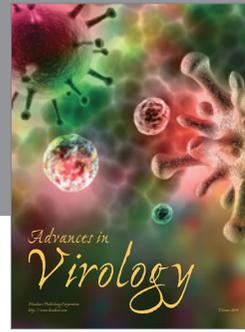
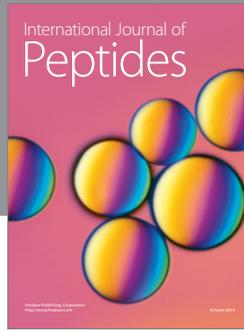
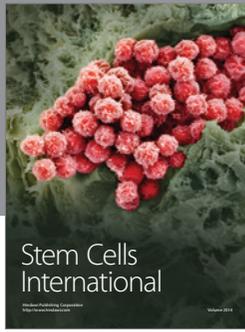
Acknowledgment

This work is partially supported by a China Scholarship Council (CSC) graduate studentship.

References

- [1] M. Archetti, "Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code," *Journal of Molecular Evolution*, vol. 59, no. 2, pp. 258–266, 2004.
- [2] L. Duret and D. Mouchiroud, "Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 8, pp. 4482–4487, 1999.
- [3] P. Tao, L. Dai, M. Luo, F. Tang, P. Tien, and Z. Pan, "Analysis of synonymous codon usage in classical swine fever virus," *Virus Genes*, vol. 38, no. 1, pp. 104–112, 2009.
- [4] E. N. Moriyama and J. R. Powell, "Codon usage bias and tRNA abundance in *Drosophila*," *Journal of Molecular Evolution*, vol. 45, no. 5, pp. 514–523, 1997.
- [5] G. P. Holmquist and J. Filipinski, "Organization of mutations along the genome: a prime determinant of genome evolution," *Trends in Ecology and Evolution*, vol. 9, no. 2, pp. 65–69, 1994.
- [6] X. Liu, C. Wu, and A. Y.-H. Chen, "Codon usage bias and recombination events for neuraminidase and hemagglutinin genes in Chinese isolates of influenza A virus subtype H9N2," *Archives of Virology*, vol. 155, no. 5, pp. 685–693, 2010.
- [7] Y. Zhang, Y. Liu, W. Liu et al., "Analysis of synonymous codon usage in Hepatitis A virus," *Virology Journal*, vol. 8, p. 174, 2011.
- [8] T. Zhou, W. Gu, J. Ma, X. Sun, and Z. Lu, "Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses," *BioSystems*, vol. 81, no. 1, pp. 77–86, 2005.
- [9] E. H. M. Wong, D. K. Smith, R. Rabadan, M. Peiris, and L. L. M. Poon, "Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus," *BMC Evolutionary Biology*, vol. 10, no. 1, article 253, 2010.
- [10] S. Kryazhimskiy, G. A. Bazykin, and J. Dushoff, "Natural selection for nucleotide usage at synonymous and nonsynonymous sites in influenza A virus genes," *Journal of Virology*, vol. 82, no. 10, pp. 4938–4945, 2008.
- [11] X. Gong, S. Fan, Z. Cui, and X. Li, "The CpG suppression of polymerase segments and its impact on codon usage bias in H1N1 influenza virus," *Acta Biophysica Sinica*, vol. 27, pp. 537–544, 2011.
- [12] N. Goni, A. Iriarte, V. Comas et al., "Pandemic influenza A virus codon usage revisited: biases, adaptation and implications for vaccine strain development," *Virology Journal*, vol. 9, p. 263, 2012.
- [13] K. Fancher and W. Hu, "Codon bias of influenza A viruses and their hosts," *American Journal of Molecular Biology*, vol. 1, pp. 174–182, 2011.
- [14] S. Katoh, "MAFFT multiple sequence alignment software version 7: improvements in performance and usability," *Molecular Biology and Evolution*, vol. 30, no. 4, pp. 772–780, 2013.
- [15] K. Katoh, K. Misawa, K.-I. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, 2002.
- [16] P. M. Sharp and W.-H. Li, "Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for "rare" codons," *Nucleic Acids Research*, vol. 14, no. 19, pp. 7737–7749, 1986.

- [17] F. Wright, "The "effective number of codons" used in a gene," *Gene*, vol. 87, pp. 23–29, 1990.
- [18] M. Greenacre and O. Nenadic, "Ca: a package for computation and visualization of simple, multiple and joint correspondence analysis," 2012, <http://www.carme-n.org/>.
- [19] R Development Core Team, *R: A Language and Environment For Statistical Computing*, Vienna, Austria, R Foundation for Statistical Computing, Vienna, Austria, 2011, <http://www.R-project.org>.
- [20] S. P. Blomberg, T. Garland Jr., and A. R. Ives, "Testing for phylogenetic signal in comparative data: behavioral traits are more labile," *Evolution*, vol. 57, no. 4, pp. 717–745, 2003.
- [21] R. P. Freckleton, P. H. Harvey, and M. Pagel, "Phylogenetic analysis and comparative data: a test and review of evidence," *American Naturalist*, vol. 160, no. 6, pp. 712–726, 2002.
- [22] M. Pagel, "Inferring the historical patterns of biological evolution," *Nature*, vol. 401, no. 6756, pp. 877–884, 1999.
- [23] L. Revell, "phytools: an R package for phylogenetic comparative biology (and other things)," *Methods in Ecology and Evolution*, vol. 3, pp. 217–223, 2012.
- [24] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree 2—approximately maximum-likelihood trees for large alignments," *PLoS One*, vol. 5, no. 3, Article ID e9490, 2010.
- [25] E. Paradis, J. Claude, and K. Strimmer, "APE: analyses of phylogenetics and evolution in R language," *Bioinformatics*, vol. 20, no. 2, pp. 289–290, 2004.
- [26] T. Britton, B. Oxelman, A. Vinnersten, and K. Bremer, "Phylogenetic dating with confidence intervals using mean path lengths," *Molecular Phylogenetics and Evolution*, vol. 24, no. 1, pp. 58–65, 2002.
- [27] Q. Liu and Q. Xue, "Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species," *Journal of Genetics*, vol. 84, no. 1, pp. 55–62, 2005.
- [28] A. Kawabe and N. T. Miyashita, "Patterns of codon usage bias in three dicot and four monocot plant species," *Genes and Genetic Systems*, vol. 78, no. 5, pp. 343–352, 2003.



Hindawi

Submit your manuscripts at
<http://www.hindawi.com>

