# Identifying sites under selection in influenza hemagglutinin

Austin G. Meyer[1,2,3,*], Claus O. Wilke[1,2]

**1 Department of Integrative Biology, Institute for Cellular and Molecular Biology, and Center for Computational Biology and Bioinformatics. The University of Texas at Austin, Austin, TX 78712, USA.**

**2 Department of Molecular Biosciences, Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, TX 78712, USA.**

**3 School of Medicine, Texas Tech University Health Sciences Center, Lubbock, TX 79430, USA.**

∗ **E-mail: austin.meyer@utexas.edu**

## Abstract

Influenza hemagglutinin is among the most studied proteins in all of viral biology. It is both the most variable gene in flu and the protein most responsible for the seasonal re-infection cycle of the human population. There have been dozens of attempts, utilizing as many different methodologies, to identify the sites that are critical for hemagglutinin's seasonal escape from the host immune system. Most of these techniques use some type of sequence analysis to identify sites that are more variable than one would expect from neutral amino acid substitutions. They often then make the jump to assume highly variable sites are under strong host immune pressure. However, since hemagglutinin is most often analyzed as a test data set for new methodologies in molecular evolution, few investigators try to connect sequence variability to actual immune epitope data. Moreover, in the last decade there has been no attempt to systematically re-analyze flu despite a ten-fold growth in available data and the crystallization of well-established molecular evolutionary techniques. Furthermore, there are a number technical complexities in handling hemagglutinin sequences like ensuring clean sequences and alignments, accurate phylogenies, and unifying site numbering between crystal structures, immature and mature proteins, and DNA sequences. For hemagglutinin H3, we have re-analyzed all currently available sequences and curated all experimental immune epitope data. We find that epitope sites are enriched for sites under positive selection. In addition, we find there are a large number of sites that are under diversifying selection that have no experimental justification for being under immune pressure; likewise there are a large number of epitope sites that are not diversifying selection.

# Author Summary

# Introduction

# Materials and Methods

# Results

## Subsection 1

### SubSubsection 1.1

## Subsection 2

# Discussion

# Acknowledgments

# References

# References

1. Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, et al. (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459: 1122–1126.

# Figure Legends

# Tables