Elucidating Influenza host-pathogen interactions through data integration and analysis utilizing the BioHealthBase BRC.

Burke Squires, Feng Luo, Marc Gillespie*, Peter D'Eustachio*, Carey Gire†, Kevin Biersack† and Richard H. Scheuermann
Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX, 75390-9072,
*Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724 †Northrop Grumman Information Technology, Rockville, MD, 20850.

BioHealthBase BRC

Introduction

The primary mission of the BioHealthBase Bioinformatics Resource Centers (BRCs) for Biodefense and Emerging/Re-emerging Infectious Diseases is to assist *Influenza virus (A, B, C), Francisella tularensis, Mycobacterium tuberculosis* researchers in their development of vaccines, therapeutics, and diagnostics. The BRCs, contracted through the National Institute of Allergy and Infectious Disease's (NIAID) Division of Microbiology and Infectious Diseases (DMID), will provide both central repositories for a wide variety of scientific data on these pathogenic microorganisms and a platform for software tools that support investigator-driven data analysis. A description of the NIAID BRC program can be found at: http://www.niaid.nih.gov/dmid/genomes/brc/default.htm

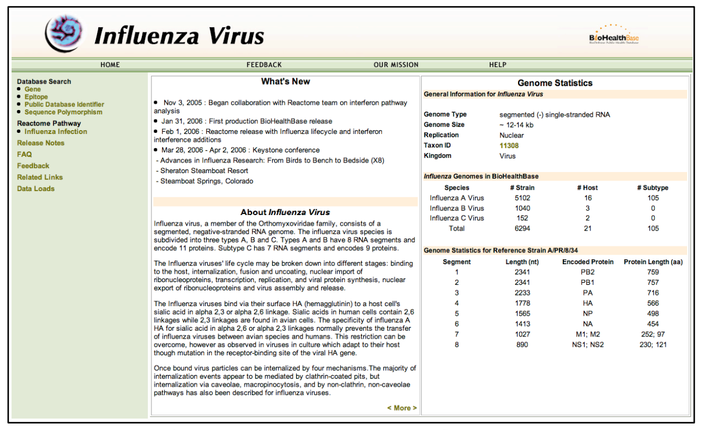

Figure 1. BioHealthBase BRC *Influenza* page.

Current Features

- Integrated data sets from NCBI, UniProt, Pfam, BioCyc, and other sources
- Web-based data-mining and visualization tools
- Sequence data repository for several strains
- Structural features and functional annotations for gene and protein sequences
- Metabolic and signaling pathway annotation
- Host-pathogen interaction models

Influenza Specific Features

- Influenza sequence alignments and polymorphism frequencies
- Creation of consensus based on sequence multi-alignments
- MHC class I epitope prediction for *Influenza* using NetCTL
- Influenza life cycle pathways and host-pathogen interactions in the Reactome database

Collaborators Welcome

We are currently seeking *Influenza* researchers who would like to work with our team to develop the BioHealthBase application. If you are interested in providing data for analysis, feedback on existing functionality, or recommendations for new functionality, please contact us at feedback@biohealthbase.org.

Reactome Pathways

Introduction

The Reactome project is a collaboration among Cold Spring Harbor Laboratory, The European Bioinformatics Institute, and The Gene Ontology Consortium to develop a curated resource of core pathways and reactions in human biology.

Figure 2. *Influenza A virus* Reactome life cycle outline.

Current and Future Work

Our initial contributions to the Reactome project includes the *Influenza A virus* life cycle stages and their subcomponents. Our immediate priority is to flesh out the *Influenza* pathways that directly relate to the known anti-viral drugs, including amantadine and rimantadine. Future work includes efforts to map out host-pathogen interactions between *Influenza virus* and a human host. Our first host pathways, RIG-I and TLR3, will be released in the next Reactome release (April/May 2006) with pathways directly interacting to follow in the future.

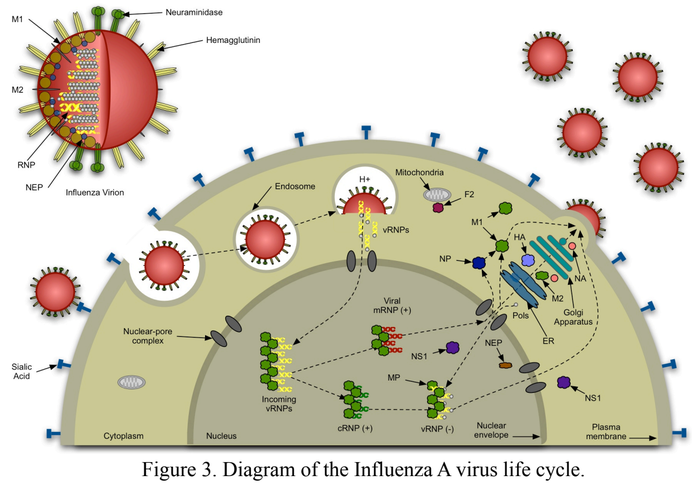

Figure 3. Diagram of the Influenza A virus life cycle.

Influenza Gene Analysis

Introduction

Influenza virus characterization to date has focused on serotype information reflecting the hemagglutinin and neuraminidase proteins expressed on the surface of the *Influenza* virion. Resources and science have limited this characterization. We propose a more complete characterization of *Influenza virus* strains based on the clades of individual proteins.

Methods

Our analysis consisted of aligning thousands of full length *Influenza* protein sequences gathered from the NCBI's *Influenza* resource using the MUSCLE program. Following the multiple sequence alignment trees were constructed using the clustalw software.

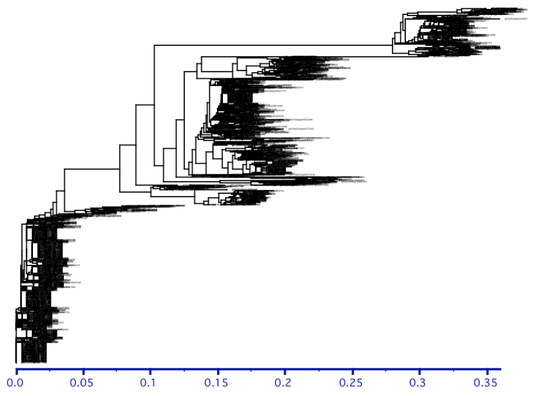

Figure 4. Influenza A virus NS1 full length protein phylogenetic tree using clustalw, following MUSCLE alignment..

A clade classification threshold was determined by single linkage clustering using Cluster3 following distance matrices calculations by the PHYLIP software protdist. We determined a threshold of 0.97 as the threshold for the hemagglutinin clades. The BLASTCLUST implementation of single linkage clustering was then used to classify the confirm the known HA and NA clades as well as determine clade for the remaining proteins.

Influenza Protein Clades (with 2 or more proteins)

Segment 1 - PB2 - 1 Clade
Segment 2 - PB1 - 1 Clade
Segment 3 - PA - 1 Clade
Segment 4 - HA - 16 Clades
Segment 5 - NP - 1 Clade
Segment 6 - NA - 9 Clades
Segment 7 - M1 - 5 Clades
M2 - 3 Clades
Segment 8 - NS1 - 6 Clades
NEP - 6 Clades

Conclusion

In conclusion, we propose a full characterization of *Influenza virus* strains based upon the clades of all 10 proteins. For example an avian *Influenza A virus* strain A/Hong Kong/156/97 might become a member of the A(1.1.1.5.1.1.5.3.5.6) superclade based on the following legend: Strain(PB2. PB1. PA. HA. NP. NA. M1. M2. NS1. NEP) ordered by segment.