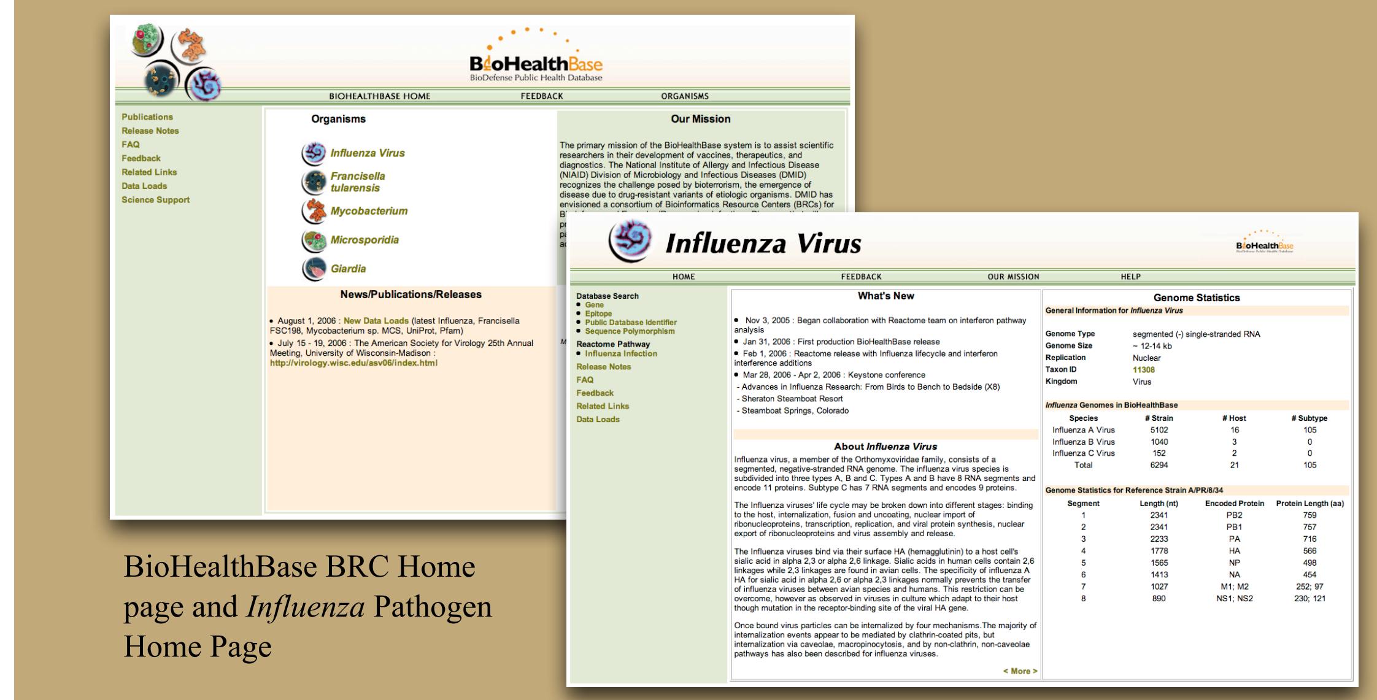
Elucidating *Influenza* host-pathogen interactions through data integration and analysis utilizing the BioHealthBase BRC.

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

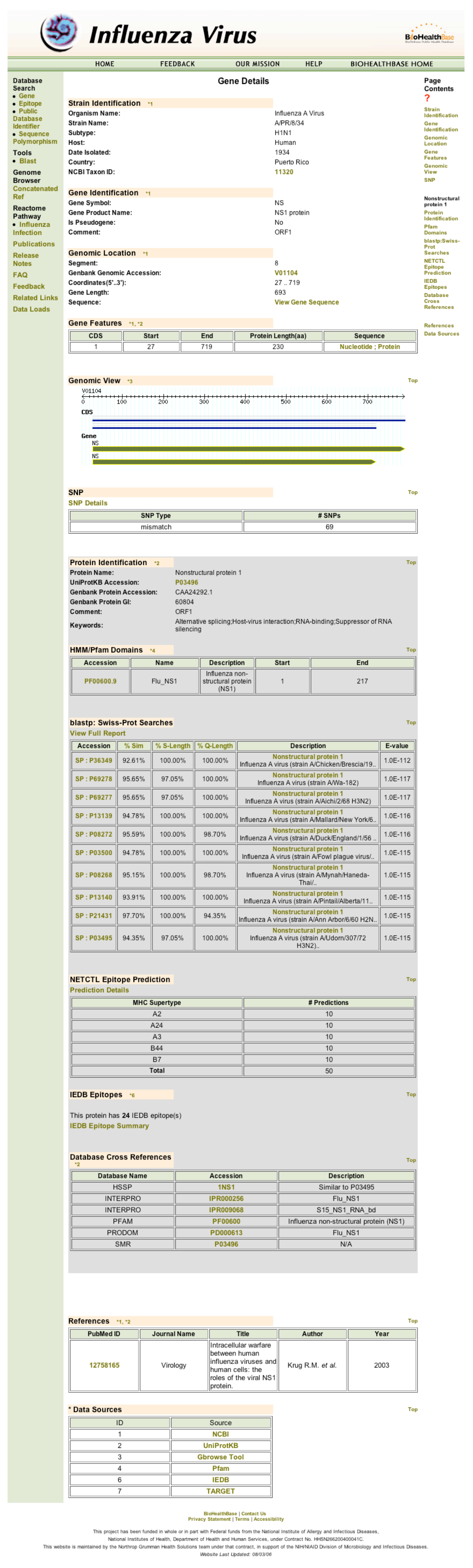


BioHealthBase BRC Home page and *Influenza* Pathogen Home Page

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* and *Microsporidia* researchers in their development of vaccines, therapeutics, and diagnostics. BioHealthBase BRC, one of eight BRCs, is contracted through the National Institute of Allergy and Infectious Disease's (NIAID) Division of Microbiology and Infectious Diseases (DMID) and provides 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

BioHealthBase BRC version 2.0 was released in August 2006 following the initial release of version 1.0 in January. Intermittent releases occur monthly focusing primarily on updating data from the various repositories and running pre-computed analysis on all data.



This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN266200400041C.
This website is maintained by the Northrop Grumman Health Solutions team under that contract, in support of the NIH/NIAID Division of Microbiology and Infectious Diseases.
Website Last Updated: 08/03/06

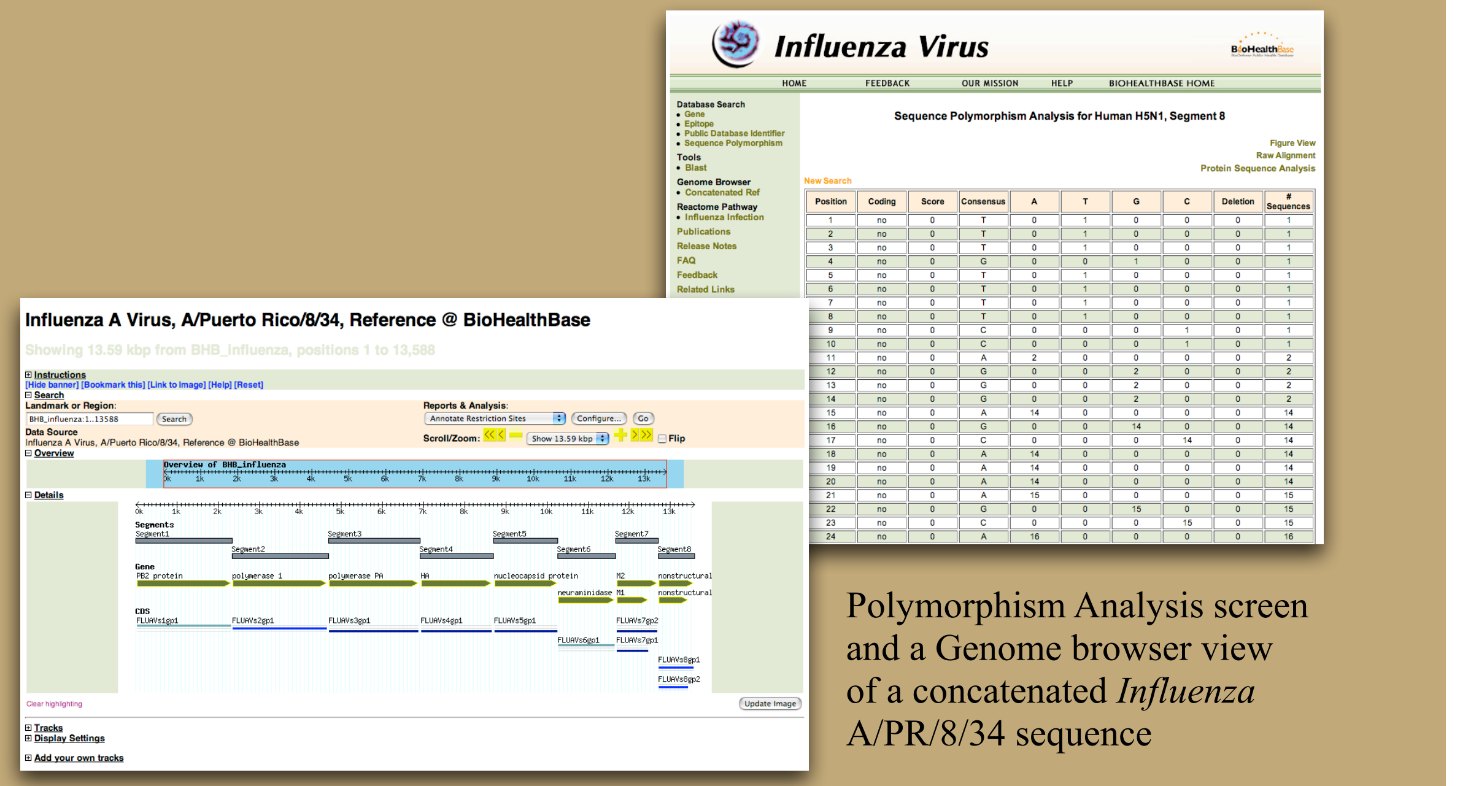
Figure 3. Gene details page for Influenza A virus A/PR/8/34 NS1 gene. Sequence features include (top-down) gene identification, location, genome view, SNP, protein information, protein domains, blast, NETCTL epitope predictions, IEDB epitopes, database cross references, references and external data sources.

Current Features

- Integrated data sets from NCBI, UniProt, Pfam, BioCyc, IEDB and other sources
- Web-based data-mining and visualization tools
- Structural features and functional annotations for gene and protein sequences
- Metabolic and signaling pathway annotation
- Host-pathogen interaction models
- Genome browser (with multiple annotation tracks)
- User-interactive blast (blastn, blastp, blastx)
- Enhanced Sequence Polymorphism search functionality (Protein Sequence Analysis)
- Additional Gene Details data display

Influenza Specific Features

- Influenza sequence alignments and polymorphism frequencies
- Creation of consensus sequence 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



Polymorphism Analysis screen and a Genome browser view of a concatenated *Influenza* A/PR/8/34 sequence

Reactome Pathways

Introduction

BioHealthBase BRC has a strong interest in understanding Host-Pathogen Interactions (H-PI). To this end we have worked with the Reactome project to add pathways for the *Influenza* virus and the corresponding human host pathways. 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.

Recent Work

Our contributions to the Reactome project follow a four stage plan as enumerated below. We have completed the initial work of the first two stages and are currently working on stage three.

Stages

1. Develop a high level framework for the *Influenza A* virus (pathogen) life cycle.
2. Develop the detailed host response pathways.
3. Develop the detailed *Influenza A* virus pathways.
4. Develop the host-pathogen linkages.

Our work to date includes the development of the high level framework of the *Influenza A* virus pathways in Stage 1, as well as the addition of the toll-like receptor 3 (TLR3) and the retinoic acid inducible gene I (RIG-I) host response pathways in Stage 2.

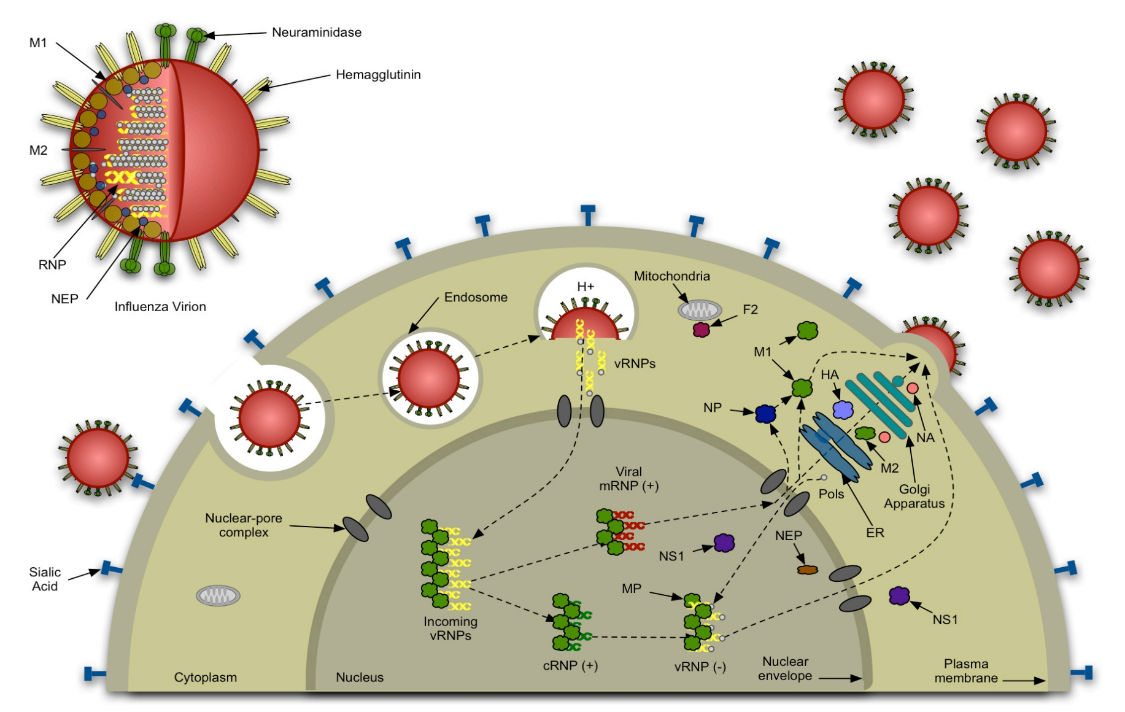
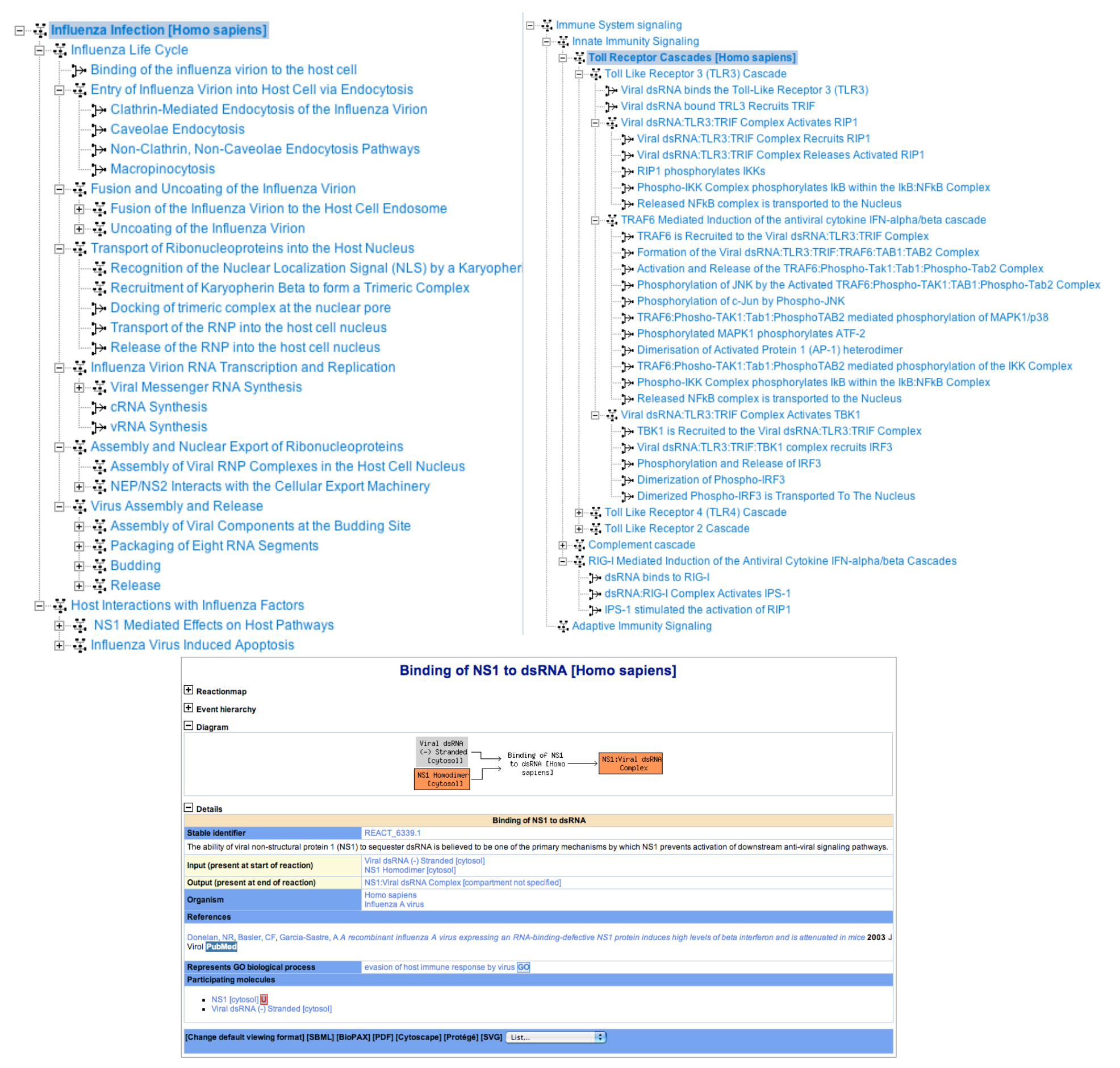


Figure 1. A diagram of the *Influenza* A virus life cycle accompanies and summarizes our efforts to add *Influenza* virus pathogen pathways to the Reactome database project.


Influenza Infection [Homo sapiens]
Influenza Life Cycle
Innate Immunity Signaling
Toll Receptor Cascades [Homo sapiens]
Binding of NS1 to dsRNA [Homo sapiens]

Donelan, NR, Basler, CF, Garcia-Sastre, A *A recombinant influenza A virus expressing an RNA-binding-defective NS1 protein induces high levels of beta interferon and is attenuated in mice* 2003 J Virol PubMed

Represents GO biological process: evasion of host immune response by virus GO

Participating molecules
- NS1 [cytosol]
- Viral dsRNA (-) Stranded [cytosol]

[Change default viewing format] [SBML] [BioPAX] [PDF] [Cytoscape] [Protégé] [SVG]

Figure 2. Our H-PI contributions to the Reactome database including an high level framework for the *Influenza* A virus life cycle (left) and the host response pathways of TLR3 and RIG-I (right). The bottom image shows the details of each step in the reactome pathways including molecules involved, a brief description, references, and Gene Ontology terms.

Current and Future Work

Building upon our previous work we begin Stage 3 with the development of some of the detailed pathways of the *Influenza A* virus. Our immediate focus is to develop detailed *Influenza A* virus pathways. The detailed pathways include:

1. Non-structural 1 (NS1) protein
2. Neurminidase (NA) protein pathways
3. Martix protein 2 (M2) ion channel

The three pathways above were selected for their important for providing host immune response countermeasures (NS1) or because of their importance as anti-viral targets. Upon completion of these detailed pathways, we will be start our work in Stage 4 as we link the NS1 pathways to the existing host response pathways. At the same time we will continue our work developing other detailed *Influenza A* virus pathways.

Influenza Protein 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. An example of a this is the avian *Influenza* designation of H5N1. Resources and science have limited this characterization. We propose a more complete characterization of *Influenza* virus strains based on the clades of individual proteins. Our analysis includes an objective way of determining the number of clades resulting from single-linkage cluster analysis utilizing a plateau characterization.

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. The BLASTCLUST implementation of single linkage clustering was then used to cluster the *Influenza* proteins. Sequential sequecne similatiry cutoffs were computed and plotted resulting in a characteristic plateau of the number of clades. The plateaus demonstrate clade that are immune to minute changes in sequence similarity.

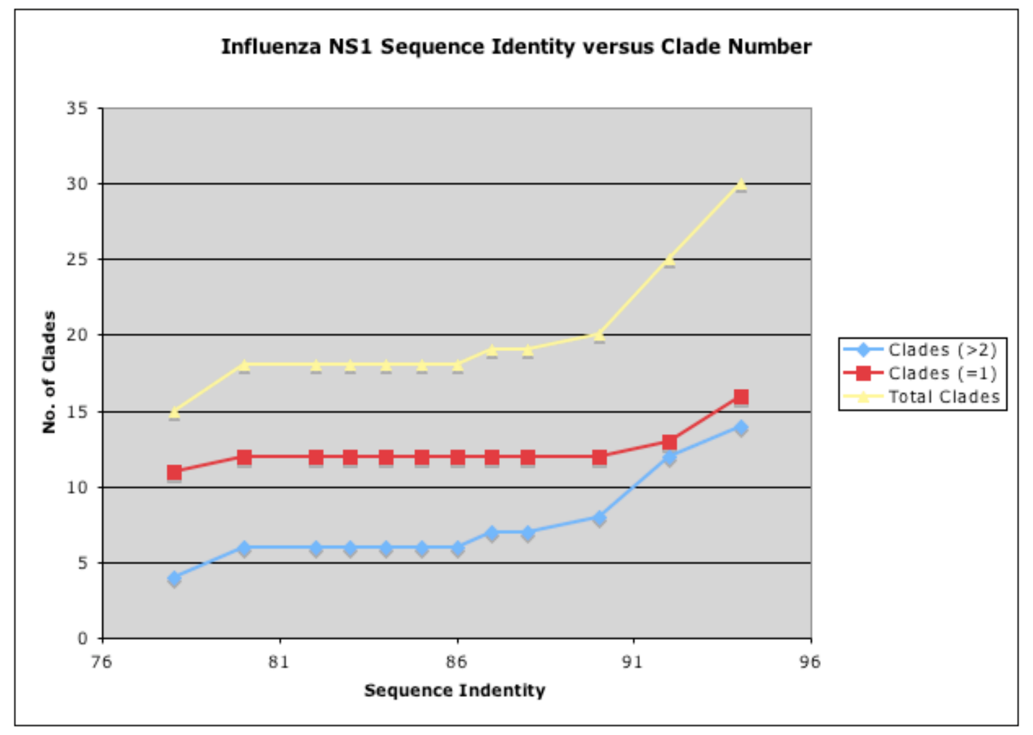


Figure 4. A graph of sequence identity threshold versus the number of resulting clades using single-linkage clustering through the BLASTCLUST program. The plateau shown between sequence identity score 81 through 86 demonstrate that the resulting number of clades are immune to small changes in sequence identity.

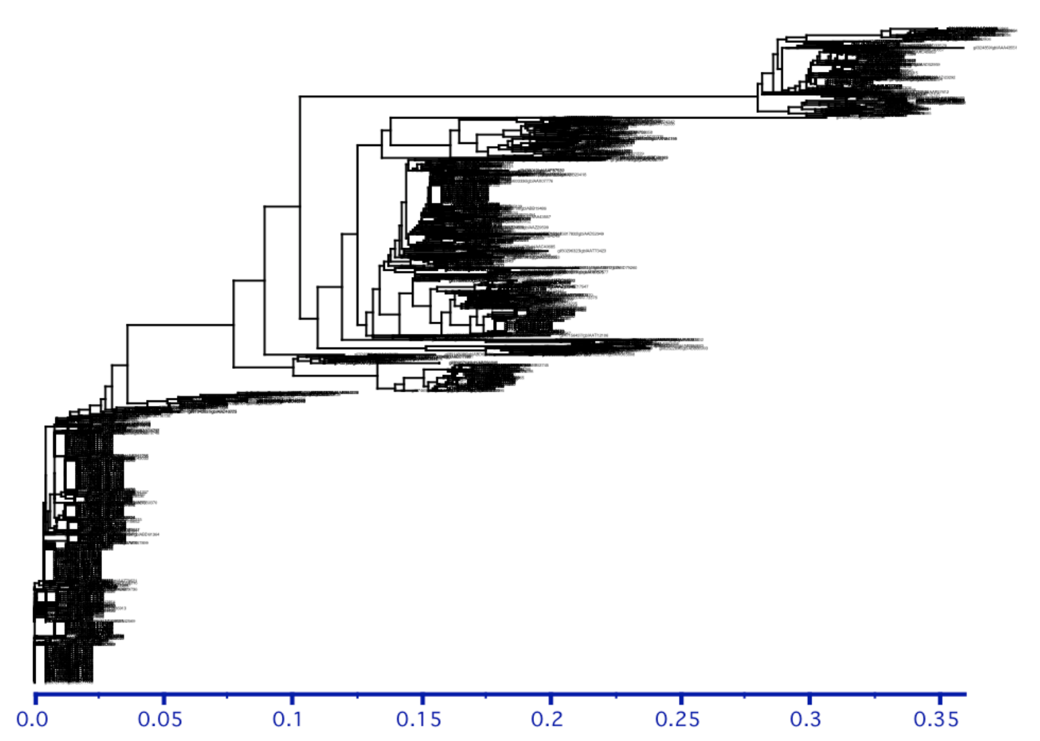


Figure 5. *Influenza* A virus NS1 full length protein sequences aligned using the MUSCL program and subsequently order as a phylogenetic tree using CLUSTALW. The results of the single-linkage clustering determined there are 6 NS1 clades in the *Influenza* samples.

Influenza Protein Clades

(with 2 or more proteins)

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

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.