ABSTRACT

Title of Document: THE VIRAL GENOMICS REVOLUTION: BIG DATA APPROACHES TO BASIC VIRAL RESEARCH, SURVEILLANCE, AND VACCINE DEVELOPMENT

Seth Schobel, Doctor of Philosophy, 2015

Directed By: Professor Michael P Cummings, Department of Biology

Since the decoding of the first RNA virus in 1976, the field of viral genomics has exploded, first through the use of Sanger sequencing technologies and later with the use next-generation sequencing approaches. With the development of these sequencing technologies, viral genomics has entered an era of big data. New challenges for analyzing these data are now apparent. Here, we describe novel methods to extend the current capabilities of viral comparative genomics. Through the use of antigenic distancing techniques, we have examined the relationship between the antigenic phenotype and the genetic content of influenza virus to establish a more systematic approach to viral surveillance and vaccine selection. Distancing of Antigenicity by Sequence-based Hierarchical Clustering (DASH) was developed and used to perform a retrospective analysis of 22 influenza seasons. Our methods produced vaccine candidates identical to or with a high concordance of antigenic similarity with those selected by the WHO. In a second effort, we have

developed VirComp and OrionPlot: two independent yet related tools. These tools first generate gene-based genome constellations, or genotypes, of viral genomes, and second create visualizations of the resultant genome constellations. VirComp utilizes sequence-clustering techniques to infer genome constellations and prepares genome constellation data matrices for visualization with OrionPlot. OrionPlot is a java application for tailoring genome constellation figures for publication. OrionPlot allows for color selection of gene cluster assignments, customized box sizes to enable the visualization of gene comparisons based on sequence length, and label coloring. We have provided five analyses designed as vignettes to illustrate the utility of our tools for performing viral comparative genomic analyses. Study three focused on the analysis of respiratory syncytial virus (RSV) genomes circulating during the 2012-2013 RSV season. We discovered a correlation between a recent tandem duplication within the G gene of RSV-A and a decrease in severity of infection. Our data suggests that this duplication is associated with a higher infection rate in female infants than is generally observed. Through these studies, we have extended the state of the art of genotype analysis, phenotype/genotype studies and established correlations between clinical metadata and RSV sequence data.

THE VIRAL GENOMICS REVOLUTION: BIG DATA APPROACHES TO BASIC VIRAL RESEARCH, SURVEILLANCE, AND VACCINE DEVELOPMENT

By

Seth Adam Micah Schobel

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Professor Michael Cummings, Chair
Professor Najib El-Sayed
Professor John Glass
Professor Sridhar Hannenhali
Professor Steven Mount

Preface

This work was supported in part by several awards as specified below. Study One was supported by award HHSO100201000061C, The Biomedical Advanced Research and Development Authority (BARDA), within the Office of the Assistant Secretary for Preparedness and Response in the U.S. Department of Health and Human Services (HHS) and funds from the Novartis Foundation (NF) and Synthetic Genomics Vaccines, Inc (SGVI). Study Two was supported in part by the National Institute of Allergy and Infectious Diseases, the National Institutes of Health (NIAID/NIH) Genomic Centers for Infectious Diseases (GCID) program, U19-AI-110819 and NIAID Genomic Sequencing Centers for Infectious Diseases (GSCID) program, HHSN272200900007C. Study Three was supported by three awards. The clinical sample and data collection for this study was supported by NIAID grant AI U19-AI-095277 and a Vanderbilt Institute for Clinical and Translational Research grant, UL1 TR000445, from NCATS/NIH. The sequencing work was generously supported by the NIAID/NIH Genomic Centers for Infectious Diseases (GCID) program, U19-AI-110819.

The opinions expressed herein are solely those of the authors and do not necessarily represent the views of the granting agencies, BARDA, HHS, NIAID, NIH, nor our corporate sponsors, NF or SGVI.

Dedication

I dedicate this work to my loving and devoted husband, Robert, and our two wonderful children, Zoe and Asher. Robert, without your support and dedication I would not be able to stand so high and reach so far. Zoe and Asher, you inspire me everyday to work to change your world for the better and to be a responsible custodian. I hope to always stand as a good example to each of you and help you to reach for the stars.

Acknowledgements

This work would not be possible without the support of a vast network of colleagues, friends and family. I would like to acknowledge these individuals for their efforts to support me in this process.

I would first like to thank my colleagues at the J. Craig Venter Institute (JCVI) for their dedication to my education and support of my research. I thank Karen Nelson for allowing me to pursue graduate research at JCVI and always showing interest in my development as a young scientist. I would also like to thank the members of the Viral Genomics and Viral Informatics groups at JCVI: Susmita Shrivastava, Paolo Amedeo, Vishal Thovari, Shiliang Wang, Brian Bishop, Danny Katzel, Neha Gupta, Reed Shabman, Kari Dilly, Jyoti Shankar, and Yi Tan. I would especially like to thank four individuals at JCVI who helped beyond measure: Kelvin Li, Karla Stucker, Suman Das, and Timothy Stockwell. Kelvin I thank for his dedication to accuracy and precision and his willingness to explain statistical concepts. I thank Karla for her words of encouragement, enthusiasm for science, and her constancy throughout this process. Suman I thank for his willingness to take a chance on me, for sharing his data, and his mentoring as a scientist. I thank Tim for countless time spent mentoring. I also thank him constant critiques, novel insights, and new perspectives on my work.

I wish to thank the members of my dissertation committee for their involvement in my research and willingness to shepherd me through the PhD process. Najib El-Sayed, Sridhar Hannenhalli, and Steve Mount, thank you for your service, interest, critiques and kind words. I thank my JCVI advisor, John Glass, without

whom I would still be attempting to piece together a dissertation proposal. I thank you for your willingness to involve me in your lab and for starting me down the long road of research ahead of me. I should also thank David Wentworth for his efforts on my committee in the past. For his mentoring as a virologist and Influenza researcher and willingness to involve me in his projects at JCVI, I will be forever grateful.

I wish to separately acknowledge my dissertation committee chair, Michael Cummings. I cannot express the extent of my gratitude for his involvement in my graduate education. Michael's calm consistency has helped me through this process in more ways than one. From advice on preparing for qualifying exams, to navigating disappointments and pitfalls, to keeping me focused and on schedule, Michael has been there and encouraged me though it all. Perhaps the best decision I have made while pursuing my PhD was asking Michael to become my advisor. Thank you, I am most grateful.

I wish to thank the many members of my family whom have assisted in the process of completing this dissertation. To my mom and dad, Penny and Steve Schobel, thank you for the encouragement and support. Thank you for the countless hours of childcare and for allowing me to write for countless weekends in your home away from the distractions of family life with small children. I also wish to thank my brother, sisters and their spouses, Brian Schobel, Lisa Porter, Sara Burke, Maggie Schobel and Eli Burke. Thank you for helping with the kids while I wrote. I also wish to thank my nieces and nephews for

I wish to thank four friends who have encouraged and supported me through this process. First I would like to thank Nikki Edworthy for your help and advice

with the scientific process and for help editing this dissertation. I thank Erin Berman for countless hours on gchat discussing science and the PhD process. Our research domains in science might not overlap, but our love for science does and that has helped to keep me focused. I thank Erin for her constant interest and kind words of encouragement. Joshua Boyle I thank for his willingness to house me during my writing mini-sabbatical. I also thank Josh for all the runs that provided a counterbalance to my cerebral pursuits. Lastly, I thank my dear friend Brianna Boyle for her constant support, her uplifting encouragement, and most importantly her friendship. Thank you Brianna.

Finally, I wish to acknowledge my husband Robert Schobel-McHugh. I thank Robert for his unwavering devotion to our children, Zoe and Asher and me. I thank Robert for countless hours of single parenthood while I performed research and wrote this dissertation. I thank Robert for lifting me up when I was down, for being a sounding board for new ideas, and for being curious about my work. I cannot express the extent of my gratitude for having you with me during this process. Thank you.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AD - Antigenic Distancing
ALRI - Acute lower respiratory tract infection
AMOVA - Analysis of molecular variance
API - Application program interface
BLAST - Basic Local Alignment Search Tool
BSS2 - Bronchiolitis Severity Score 2
BaTS - Bayesian Tip-association Significance Testing
CDS - Coding sequence
CIPRES - Cyber Infrastructure of Phylogenetic Research
DASH - Distancing of Antigenicity by Sequence-based Hierarchical Clustering
F - Fusion
FDA - Food and Drug Administration
G - Glycoprotein
GARD - Genetic Algorithm for Recombination Detection
GSI - Geneological Sorting Index
GTR-IG - General time reversible model with a nucleotide site-specific rate heterogeneity with four rate categories and invariant sites
HA - Hemagglutinin
HI - Hemagglutination inhibition assay
HLA – Human leukocyte antigens
HMM - Hidden Markov model
HPD - Highest posterior density
IAV - Influenza A virus
IBV - Influenza B virus
ILI - Influenza-like illness
INDEL - Insertion/Deletion
INSPIRE - Infant Susceptibility to Pulmonary Infections and Asthma Following RSV Exposure
IRD - Influenza Research Database
JCVI - J. Craig Venter Institute
L - Polymerase
M - Matrix
M2-1 - Transcription regulator 1
M2-2 - Transcription regulator 2
MAR - Manufacturing at-risk
MCC - Maximum clade credibility
MCMC - Markov chain Monte Carlo
MDS - Multi-dimensional scaling
ML - Maximum likelihood
MLST - Multi-locus sequence tags
MetaCats - Metadata-driven comparative analysis tool for sequences
N - Nucleocapsid
NA - Neuraminidase

NGS - Next-generation sequencing
NH - Northern Hemisphere
NIGSP - NIAID Influenza Genome Sequencing Project
NIMR - National Institute for Medical Research
NJ - Neighbor joining
NS1 - Non-structural protein 1
NS2 - Non-structural protein 2
OBO - Open Biomedical Ontologies
ORF - Open reading frame
P - Phosphoprotein
PCA - Principal component analysis
RDBMS - Relational database management system
RDRP - Recombination Detection Program
REST - Representational state transfer
RMSD - Root-mean squared deviation
RSV - Respiratory Syncytial Virus
RT - Reverse transcription
SBP - Single Breakpoint Recombination
SH - Small hydrophobic protein
SH - Southern Hemisphere
TSV - Tab separated values
URI - Upper respiratory infections
VIGOR - Viral Genome ORF Reader
VRBPAC - Vaccines and Related Biological Products Advisory Committee
ViPR - Viral Pathogen Resource
WHO - World Health Organization
mAB - Mouse antibody
tMRCA - Time to most recent common ancestor

Chapter 1: Introduction to Viral Genomics and Comparative Genomics

Viral genomics began with the sequencing of the first RNA virus in 1976 (6) soon followed by the sequencing of the first DNA virus using the Sanger method of sequencing in 1977 (7). Both of these viruses were relatively small bacteriophages, MS2 and Φ X174 respectively (6, 7). Due to its amenability for use as a high-throughput technique, Sanger's method of DNA sequencing (8), using dideoxy chain-terminating nucleotide analogues, was responsible for the great genomics revolution of the 1990's and 2000's. Starting with radiolabeling and later with florescent labeling, this technology led to the sequencing of dozens of eukaryotic genomes, hundreds of bacterial genomes, and thousands of viral genomes. Recently, Sanger sequencing has been supplanted by several next-generation sequencing techniques (NGS) that use massively parallel sequencing approaches to greatly accelerate the speed at which DNA sequences can be produced (Figure 1).

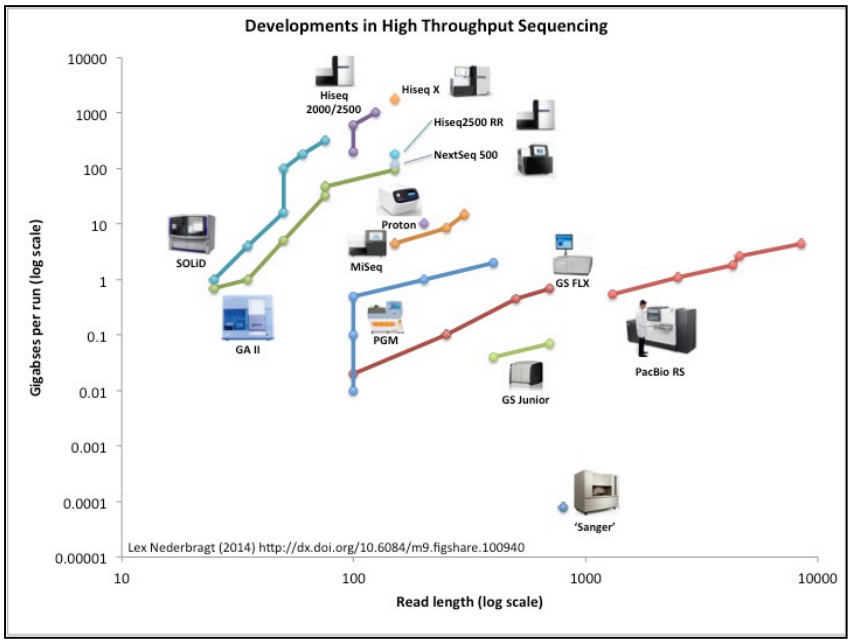


Figure 1 Comparison of Sanger and Next-Generation Sequencing Platforms (3).

Viral genomics has benefited significantly from these technological advances, enabling sophisticated epidemiological and evolutionary analyses and providing novel insights into the pathogenicity, ecology, and epidemiologic nature of countless human and animal pathogens.

Viral Genomics

Background

Early viral surveillance, evolution, epidemiology, and molecular pathogenesis studies typically did not rely on whole genome sequencing. Rather, small portions of viral genomes were sequenced to answer specific questions regarding the viruses in question. For example, influenza virus surveillance largely relied on the sequencing of the hemagglutinin and neuraminidase genes due to fact that these genes could be used to genotype the virus. Knowing these genetic types allowed researches to track the evolution of the virus and predict when the virus had drifted sufficiently to require a new vaccine. Similarly, respiratory syncytial virus (RSV) researchers focused their sequencing efforts on the G and F proteins, as these genes could be used to discriminate between the circulating lineages of RSV.

Whole genome approaches, as enabled by high-throughput sequencing technologies, allow a deeper understanding of viral evolutionary and epidemiological dynamics. For instance, whole genome studies provide researchers with the ability to identify reassortment in segmented or recombination events in non-segmented viruses. These types of genomic rearrangements give viruses a mechanism to increase their genetic diversity and hence increase their chances of acquiring a selective advantage. Whole genome studies also provide researchers with insights

into the epistatic interactions between mutations. These epistatic interactions could have implications for the molecular pathogenesis and epidemiology of the virus. An example of this was published by Das et al showing that through monoclonal antibody challenge epistatic mutations could be elicited in influenza virus proteins to provide both an immune escape at one site and compensatory protein stabilization at a second and even third site (9). The stabilization described could be structural or functional in nature (9). This type of research is especially important to uncover possible mutational paths available at any given point in the evolution of human pathogens.

Whole genome sequencing also enables a wide range of methods for performing viral research. These methods include generating virus stocks from infectious clones for single-stranded positive-sense virus, or from a reverse genetics system for single-stranded negative-sense and double-stranded viruses. These systems are used to generate viruses in cell culture without the need for access to natural isolates (Figure 2). Combined with synthetic technologies, viruses can

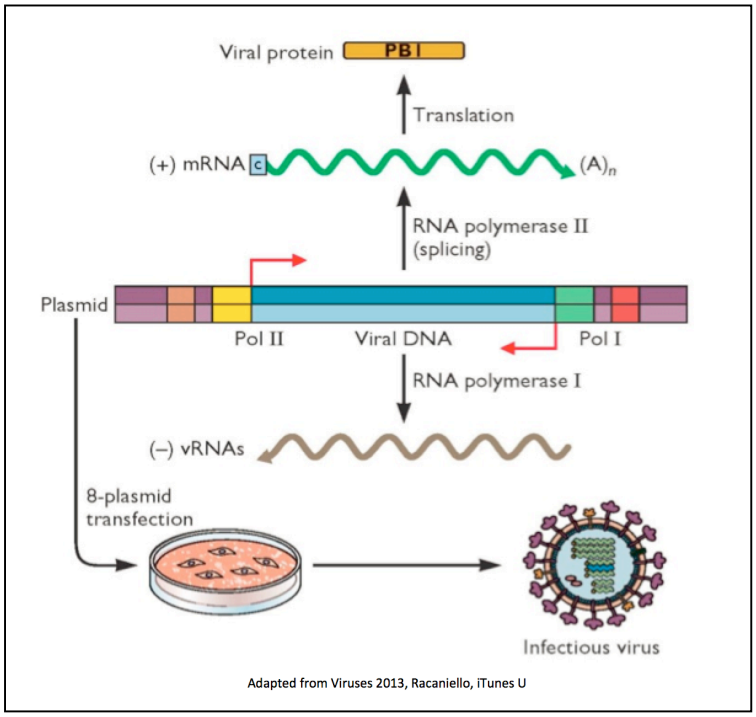


Figure 2 Generation of Infectious Virus Using a Reverse Genetics System (2).

be generated simply from the digitally stored whole genome sequence of a virus, without the need for any of the starting DNA used with traditional molecular cloning techniques. The ease of generation of viral stocks has accelerated a wide range of molecular studies, such as the antigenic testing of viruses, replacing the previous use of the hemagglutination inhibition assay to assess antigenic similarity of influenza viruses. This has the potential to positively influence the process by which influenza virus vaccine candidates are selected. Similarly, specific mutations of interest can be identified using whole genome sequencing, those mutations quickly synthesized into virus rescue systems and the viruses tested either in *in vitro* or *in vivo* testing models.

Traditional Viral Sequencing Using the Sanger Method

Until the advent of next-generation sequencing technologies in the mid to late 2000's, viral genome sequencing was performed using traditional Sanger sequencing methodologies. To prepare viral genomic materials, the viral class must first be determined. The genetic material of DNA viruses can be sequenced directly, however RNA viruses must first undergo a reverse transcription (RT) step to convert the genomic material from RNA to DNA. Sanger and next-generation sequencing can only be done on DNA. Once the source material is prepared, Sanger sequencing using specific primers can be used. Primers will differ depending on whether a sequence-dependent or sequence-independent method was used to generate source material.

Viral Sequencing Using Next-Generation Technology

Viral genomic sequencing using next-generation technology has enabled a rapid acceleration of the completion of the sequencing of viral genomes in recent years. Platforms such as Roche 454, Illumina MiSeq and HiSeq, and Life Technologies Ion Torrent are a few used to generate genome sequences in a massively parallel fashion. These technologies all produce huge amounts of sequence by sequencing in parallel small fragments of the source material. Typically, the source material is first prepared for sequencing by constructing libraries with technology-specific adaptor sequences. In the case of viral genomes, whole runs on next-generation devices for individual viruses are not necessary. Multiplexing techniques are used to combine multiple viral genomes onto a single next-generation runs. Multiplexing is also accomplished by attaching unique DNA barcode sequences (typically six to eight nucleotides in length) to the genome fragments. This allows for the genomes from single runs to be de-multiplexed and binned per genome or genomic library after sequencing is complete.

Sequence-Independent Methods

Sequence-independent viral sequencing methods are typically used for viral discovery or for when no close relatives of the target virus are available. Sequence-independent methods do not rely on the use of specific primers for the target genome. Instead, these methods use random priming, usually with tagged primers to amplify the source material prior to sequencing (10). Random priming techniques will amplify both the target genome and any other genetic contamination present in the sample, therefore, enrichment techniques aimed at concentrating viral genetic

material and excluding host sequences are recommended (11). Using Sanger methodology, sequencing is conducted with primers specific to the tag sequences on the random primers (10). Using next-generation approaches, technology specific libraries are constructed with the enriched and amplified products (11).

Sequence-Dependent Methods

Sequence-dependent viral genomic approaches require that close references of the target genomes are available in the public sequence repositories. Target genomes are amplified prior to sequencing using primers specific to the genus or species of virus of interest. Typically, genomes must contain 90% sequence identity for PCR primers to be designed and used to amplify target genomes (12). Given adequate reference sequences, primers can be designed with automated computational pipelines and such computer generated primers have exceptionally high reaction success rates (12). Once genomes are amplified, usually using several PCR reactions (one or more reaction per genomic segment for segmented viruses, or several for non-segmented genomes), sequencing can be performed. For Sanger technologies, PCR products are sequenced directly, whereas with next-generation approaches the PCR products are fragmented and technology-specific sequencing libraries are produced prior to sequencing.

DNA Versus RNA Viruses

Viruses come in a variety of classes according to the Baltimore classification system. The most notable difference between the classes, as it applies to sequencing, is that there exist both RNA and DNA viruses. DNA viruses can be largely

sequenced using the same basic methods as any higher order organism; however, RNA viruses must first undergo a RT step to convert the RNA into DNA prior to sequencing. Several RT methods exist, most notably multiplexed reverse transcription polymerase chain reaction or mRT-PCR (13, 14). In an mRT-PCR reaction on an influenza A virus genome, all eight genomic segments can simultaneously be reverse transcribed and amplified in a single reaction tube (13). This greatly accelerates the speed at which influenza viruses can be prepared for sequencing. Not all viruses have an efficient mRT-PCR method developed for them. Rotavirus, for example, currently requires separate RT-PCR reactions for each genomic segment to adequately amplify the cDNA (11).

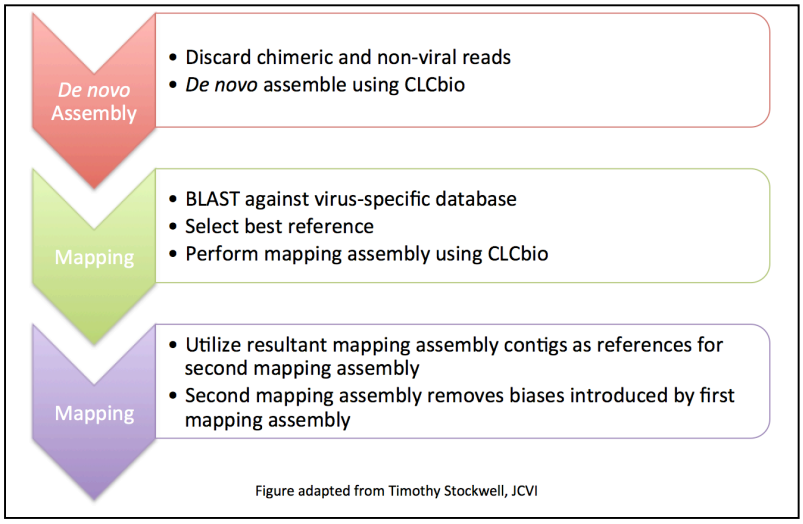


Figure 3 *De Novo*-Mapping-Mapping Assembly Method Using CLCbio (1).

Assembly

Whether sequenced using Sanger or next-generation methodologies, the next step in data preparation for viral genomic sequences is assembly. Initially, the NIAID Influenza Genome Sequencing Project (NIGSP) relied on a custom assembler called Minimus that was capable of assembling Sanger sequences with minimal overlaps (15). This was possible due to the fairly constant genomic structures in closely related influenza virus genomes. More variable

genomes require different methods for assembly. Currently, the J. Craig Venter Institute (JCVI) employs a *de novo*-mapping-mapping assembly approach to genome assembly using next-generation sequences (Figure 3) (1). In this method, genomic sequences are first assembled using CLCBio's *de novo* assembler (16). The resultant contigs are put into NCBI's Basic Local Alignment Search Tool (BLAST) (17) against a reference database of genomes presumed to be similar to the subject. The closest reference for each genomic structure is selected for a CLC mapping assembly (18). The consensus of this first round of mapping assembly is then used as a reference for a second round of CLC mapping assembly. This second mapping assembly pulls in sequences that are more distant from the original reference. This *de novo*-mapping-mapping approach has been successfully used for a variety of viral genome species (1). Drawbacks of this approach come up when dealing with genomes where no close reference sequence is available for mapping, or when the subject genome is recombined with respect to two or more reference genomes (1). Both of these scenarios require manual intervention in the assembly process, which in turn slows the pace of the assembly process (1).

Annotation

Genome annotation is the process by which genes are located and protein sequences are assigned annotation data types (such as gene symbols, protein names, go terms, or enzyme commission numbers). In higher order organisms, the discovery of open reading frames (ORFs) in nucleotide sequences is accomplished by statistically assessing each base pair in the genome in its context to determine if it is likely to be a coding or non-coding base (19, 20). This assessment can be used to

string together many so-called coding bases to form ORFs. In a second stage of annotation, the ORFs or resultant translated amino acid sequences can be used as a query in a database search with BLAST or run through hidden Markov model (HMM) search programs to assess similarity to proteins with a known function. If adequate similarity is determined, functional assignments of the aforementioned annotation data types can be made. Viral annotation requires a different approach compared to that used for other genomes due to the typically small genome sizes of many viral species. Viruses rely on densely coded genomes to transmit their genetic material and carry out the functions of viral cellular infections. Statistical approaches to ORF calling are not applicable, since very little non-coding sequence is available to build appropriate ORF calling models (21, 22). Instead, successful viral genome annotation requires a homology-based approach to ORF calling and functional annotation. Homology-based approaches require highly curated reference databases containing all the target proteins of interest needed to adequately annotate a viral species (21, 22). Annotation programs, such as the Viral Genome ORF Reader (VIGOR), utilize the homology-based approach (21, 22). VIGOR first performs a BLASTX (nucleotide query against a protein subject database) against the reference database. High-scoring hits are clustered and merged according to genome location and the best hit(s) are selected. For each hit, the start and stop codons are located. If the complete coding sequence (CDS) for a given protein is identified, it can be assigned the full annotation from the reference database for the gene that the CDS represents. If the start or stop codons are missing or are in less than ideal locations, the assignment of partial or pseudogene can be made to the CDS, indicating a lower

confidence for the structural annotation. VIGOR can identify polyproteins and mature peptides in viral genomes that then translate all or several viral proteins in a pre-processed state; these proteins are later cleaved into smaller functional viral proteins (21, 22). Approaches like VIGOR also allow for novel viral features, such as RNA editing or ribosomal slippage, to be identified and adjustments to the CDS and amino acid translations can be made accordingly (21, 22). Despite the fact that homology-based methods are powerful tools for viral genome annotation, they do require the expert curation of an appropriate reference database, a potentially time-consuming process that can necessitate consultation with experts on specific virus families.

Viral Genomic Standards

The viral genomics revolution has brought together data from various smaller communities of researchers. For this reason, special attention to viral genomic standards should be made as we expand our knowledge of the vast varieties of viral species that exist in our world. Where once community or even researcher driven standards of annotation where sufficient to understand a particular virus, now common data types across viral species can be made abstract and applicable to all viral genomic studies. A few features common to viral genomics projects amenable to standardization include viral strain names, protein names, and metadata types: such as passage medium and number for cultured viruses, host species designations, and collection methods or even collection locations. Use of common data types across viral genome projects will serve to enable access to viral genomic research across species without the need for specific domain knowledge or expertise. High-

throughput computational analysis could also benefit from such standardization. This need is apparent, for example, in the use of standardized strain names. Influenza virus strain A/California/07/2009 could be written any number of ways (such as A/CA/07/09, or A/California/7/09). To the human eye these names appear to describe the same virus; however, computers require explicit rules to be established in code to make the same associations. Similarly, passage metadata for an influenza virus may be listed as "Siat2" or "E3". A human knowledgeable of influenza passaging techniques would know "Siat2" means passaged in siat cells (a cell line derived from canine kidney cells) twice, or that "E3" means passaged in embryonated eggs three times. Again, such nomenclature requires computer code to enable automated use of such fields. This is especially true of passaging information that is not standardized, such as when researches code information using synonymous texts like "S2", "SIAT2", "Siat 2", "Siat-2" or other variants. Similarly, standards for protein naming are essential. NES and NS1 both refer to the same influenza virus protein. If one is used, the synonym should also always be listed. In short, some thought on genome standards should be given, and standards should be developed and enforced across all viral species for naming, annotation data types, and metadata in order to enable computer analysis of these data, and for the sake of accurate analysis.

Viral Comparative Genomics

Background

Preparation of viral genomic sequences is merely the first step in understanding the intricacies of viral ecology, evolution, molecular pathogenicity and epidemiology. Once a single or group of viral genomes have been completed, they

then need to be placed in context of other viruses in order to glean insights about the virus of interest. This can be done using a number of techniques and together these techniques form the basis for modern viral comparative genomics.

Public Resource Databases

Public repositories of viral sequence, phenotype and related metadata are essential resources for any comparative viral genomics study. Public databases can be categorized into two basic forms: those that simply enable data archiving and those that also allow for bioinformatics analysis. Some public repositories of note are listed in Table 1. Beyond the sequences generated from a study, the addition of context sequences is essential to perform useful analyses.

Table 1 Select Web-Accessible Public Repositories of Viral Genomic Data.

Resource	Short Name	Viral Taxa	URL
Viral Pathogen Resource (23)	ViPR	All	http://www.viprbrc.org
Viral Zone (24, 25)	ViralZone	All	http://viralzone.expasy.org
International Repository for Hepatitis B Virus Strain Data (26)	HepSeq	HBV	http://www.hpa-bioinformatics.org.uk/HepSEQ-Research/Public/Web_Front/main.php
Heptatitis C Virus Database (26)	HCV Database	HCV	http://hcv.lanl.gov
HIV Sequence Database (26)	HIV Database	HIV	http://www.hiv.lanl.gov
Influenza Research Database (27)	IRD	Influenza	http://www.fludb.org
Influenza Virus Resource (28)	IVR	Influenza	http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html
Global Initiative on Sharing Avian Influenza Data (29)	GISAID EpiFlu	Influenza	http://platform.gisaid.org
Network of Expertise on Animal Influenza	OFFLU	Influenza	http://www.offlu.net

In recent years, database resources that go beyond simple data archiving functionality have been engineered. Viral comparative resources, such as the Influenza Research Database (IRD) and Viral Pathogen Resource (ViPR), aim to be complete bioinformatics platforms that can be used to launch comparative viral

genomic studies (23, 27). These resources have had some notable successes and some drawbacks have been found. Successes include the implementation of standards for viral genome annotation and metadata types (30, 31), metadata-driven sequence analysis (32), and implementation of maximum likelihood phylogenetic analysis (23). Drawbacks include a limited set of analysis applications. For instance, ViPR does not instantiate Bayesian phylogenetic tools for use with database-derived or user-supplied data sets; although, Bayesian tools have been implemented on other web-based resources, such as Cyberinfrastructure of Phylogenetic Research (CIPRES) (33). Another limitation to a web-based approach to bioinformatics is that it limits the ability of advanced users to perform high-throughput or custom analyses.

Phylogenetics and Related Methods

Phylogenetic analysis techniques are among the most important set of methods for performing viral comparative genomics. Phylogenetics is the study of the evolutionary relatedness between species or other entities most often based on data for morphological characters or biological sequences (nucleotides or amino acid). Early algorithms for performing phylogenetic analyses were extensions of hierarchical clustering algorithms; however, with the advent of more complex evolutionary models and modern computing technologies, the collection of phylogenetic methods available today extends the capabilities of performing evolutionary studies significantly.

Due to the computational demands and availability of advanced phylogenetic methodologies, older viral comparative studies relied on methods such as neighbor joining (NJ) to perform evolutionary analyses. NJ is a deterministic algorithm that

operates on distance data to perform a heuristic search of tree topologies (34). Many have speculated on the precise nature of the NJ algorithm and the optimality criterion it aims to solve (34, 35). Given the same input data and evolutionary assumptions it will give the same resultant tree each time the program is run, although this tree may not be the most optimal solution.

With the development of more advanced algorithms and modern computer processing technology, phylogenetic methods such as maximum likelihood (ML) and Bayesian techniques have been applied to phylogenetics. ML can be coupled with advanced substitution models, such as the general time reversible model, to infer evolutionary change between sequences. NJ is a deterministic algorithm, that given the same input and evolutionary model, will arrive at the same output tree (34, 36). In contrast ML searches for the tree that best fits a particular evolutionary model that satisfies the maximum likelihood criterion via a stochastic search of tree space (37) utilizing strict or non-strict hill climbing algorithms.

ML algorithms aim to find the most likely phylogeny from a given a set of input characters. ML methods are thought to be robust to violations in the substitution model because they are able to account for hidden evolutionary changes that were likely to occur over time, at some probability as determined by the model used to make the phylogenetic inferences. Even with relatively few taxa, the space of all available tree topologies vastly dwarfs the number of trees the ML algorithms are capable of examining under reasonable time scales. Tree space is also non-uniform. Strict hill climbing algorithms are more likely to get stuck in local maxima without finding the true ML tree. This can also be true for non-strict hill climbing algorithms,

but to a lesser degree. Ultimately the ML tree is just a point estimate, with a rational statistical methodology to support the evolutionary relationships it describes. For these reasons, methods for assessing the quality of the ML tree are required. Most notably, bootstrapping is typically performed to assess the reliability of specific node branching topologies found within the tree (36). Similarly, multiple ML analyses can be run to increase confidence that the ML phylogeny has been well sampled. Together, these methods form the basis for best practices in ML phylogenetics. It should be noted that bootstrapping is also routinely performed on NJ inferred phylogenies.

Bayesian phylogenetic techniques are now frequently performed in viral phylogenetics and offer a framework for performing state-of-the-art spatiotemporal analyses of sequence data. Bayesian statistics has the advantage of using prior knowledge of a problem of interest to help inform the analysis of the problem (36). It is different from ML methods in that Bayesian methods aim to sample multiple possible phylogenetic hypotheses and use all of them to generate distributions of the various parameter statistics related to the inferred phylogenies (38). These statistics range from probabilities for specific branch length and node topologies, to mutation rates or times to most recent common ancestors. The distributions of these statistics can be used to infer credible intervals around various aspects of the phylogeny in question and thus provide confidence in the phylogenetic hypothesis. These intervals are calculated on distributions collected primarily as part of the Markov chain Monte Carlo (MCMC) algorithm (36), as opposed to after-the-fact using bootstrapping in an ML context. The difficulties of Bayesian approaches are the computational

complexity of the analyses and the selection of the priors. Computation for a Bayesian analysis can take days to months on modern computers, although some analyses can be completed in much less time (minutes to hours). It should be noted that the time to complete any particular analysis is largely dependent on the size of the data set and the complexity of the model being examined. Additionally, the selection of priors can sometimes be difficult, necessitating the rerunning of analyses to find the right parameters to use for a given data set.

Methods related to phylogentics allow researchers to measure the variability and the selective pressures seen in viral data sets. These methods include entropy analysis, which simply calculates the variability seen at specific sites within the viral genome or within the viral gene of interest (39), and dN/dS analysis, which quantifies the proportion of non-synonymous substitutions at non-synonymous sites to synonymous substitutions at synonymous sites within the coding genes of a genome (40, 41). The latter example is used to assess whether specific sites within a genome are undergoing positive, neutral (diversifying), or purifying selection. Each type of selection has a specific, measurable signature that can be use to inform downstream analysis and future research activities. For example, high numbers of positive selected sites in a protein tells us that the protein is likely to be experiencing some sort of selective pressure in its environment. This pressure could be exerted by the immune response of the host or be seen as the mutations required for a pathogen to adapt to the biological environment of a new host species, in the case of an emerging pathogen. Both examples can help direct future research, inform us about a vaccine's efficacy, or make inferences about trends in global virus surveillance.

Regardless of the methodology used, phylogenetics is an essential tool for making important inferences about viral epidemiological and pathogenic features. Examining an influenza virus hemagglutinin gene tree, for instance, can suggest to researchers that the gene is under selective pressures, presumably host immune system pressures (42). Close examination of a viral phylogeny with geographical data can reveal if local or global circulation is seen during an epidemic (43). Trees can put viral studies into an ecological context, such as identifying which bird species are involved in the spread of avian influenza virus across the North American flyways (44). This information can give researchers insights into how highly pathogenic outbreaks of influenza virus originate. These are just a few examples of the power of phylogenetics to draw important conclusions about viruses.

Phylodynamics

Phylodynamics is a relatively new concept in phylogenetics. It encompasses a set of methods related to Bayesian phylogenetics that enable researchers to look at the epidemiologic, immunologic and ecological factors that lead to specific viral phylogenies (45). The most notable of these techniques is phylogeography. Phylogeography enables researchers to examine how viruses spread across geographic spaces (46). This is especially useful in tracking emerging infections as they traverse geographic space from one host to the next, and enables researchers to make important inferences about the epidemiological characteristics of these pathogens. Phylodynamics is, in essence, a merging of important viral metadata or phenotype information with phylogenetic analysis to make additional quantifiable

inferences about viral evolution and the notable characteristics of viral pathogens (45, 47).

Genome Constellation Analysis and Recombination Analysis

Although phylogenetics tells us a great deal about the nature of viral ecology, the complexity of phylogenies makes interpretation of the results challenging. Categorical assignments of specific viral strains or viral genes can simplify these complex relationships for ease of interpretation. One such methodology of categorical assignment of viral genes is known as genome constellation analysis, sometimes referred to as genotype analysis. Constellation analysis is simply a way of grouping specific genes between viral genomes together based on some sort of objective criterion. These methods utilize simple percent identity or phylogenetic distances to establish categorical membership (48-52). Once categories are assigned, constellations between viral strains can be compared to determine a number of characteristics. These include: evidence of reassortment events in segmented viruses, recombination in non-segmented viruses, and overall genetic diversity. Combined with other viral data types, this information can reveal relationships between the genotype and the phenotype.

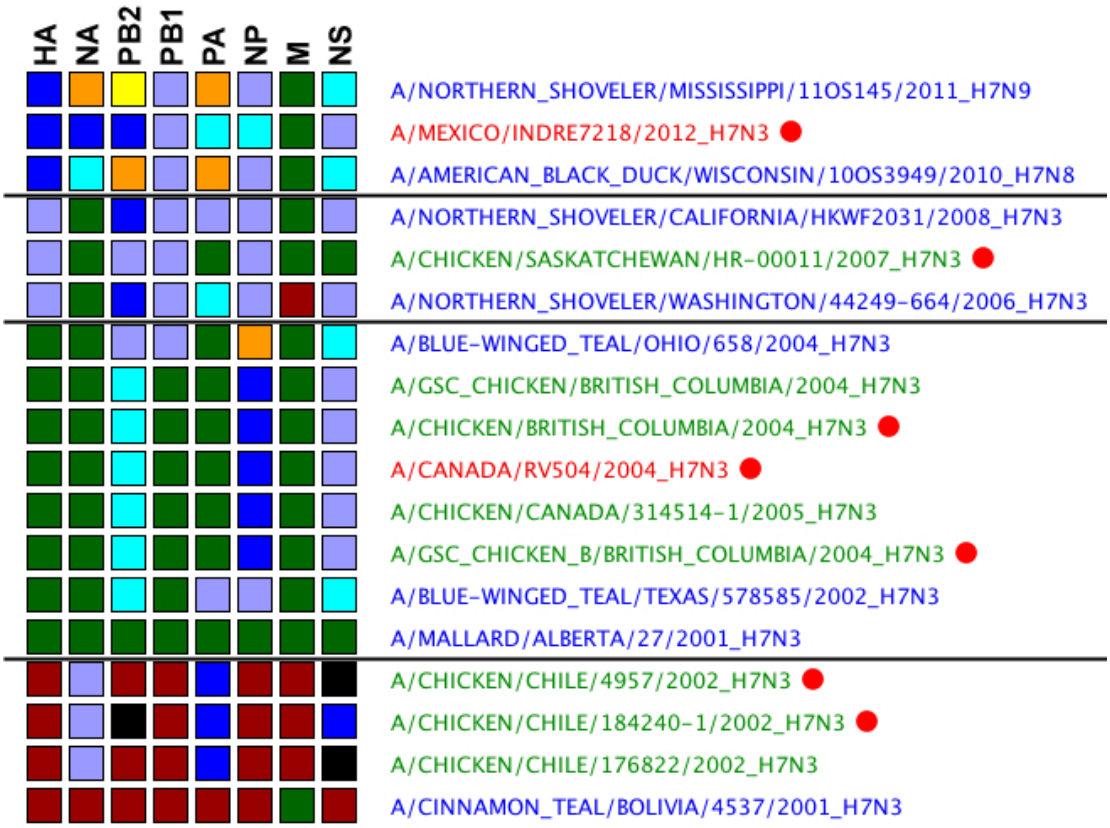


Figure 4 Example Genome Constellation Figure Depicting Avian Influenza H7 Genomes. The genome constellation analysis performs independent categorization of each gene per column. Colors for individual genes are not related to identical colors in other columns and are assigned arbitrarily by gene cluster number. Viral strains are listed per row. In this case highly pathogenic influenza strains are indicated with red dots by the name of the strain in a post analysis annotation step (4).

Although constellation analysis can reveal recombination events, other tools exist to identify probable recombination within viral data sets. These tools include the Recombination Detection Program (RDP), Single Breakpoint Recombination (SBP), and the Genetic Algorithm for Recombination Detection (GARD). RDP implements several recombination detection algorithms and provides a graphic that denotes the probable locations of the recombination events, as well as the probable donor of the recombinant subsequences (53). Similarly, SBP and GARD (both methods implemented on the datamonkey.org website), are capable of detecting the presence of recombination events and identifying the locations of the recombination

events, respectively (54). These algorithms are more statistically grounded than the detection through constellation analysis, and should be considered higher order evidence. It is important to note, however, that recombination event detection should always be confirmed, as methodological errors in sequencing can lead to the artificial detection of recombination events.

3-D Structural Analysis

Structural analysis of viral proteins is a key component of research aimed at determining the effects of mutation on viral phenotypes. 3-D structures of proteins can be compared between viruses to determine the structural changes with the potential to alter antigenic phenotype, viral attachment efficiency, or any number of other phenotypes associated with molecular pathogenicity and epidemiology. For example, the hemagglutinin protein of influenza has been reported to alter its glycosylation pattern as part of its strategy to evade host immunity. These types of changes can be modeled as 3-D structures for ease of visualization. One drawback to 3-D structural analysis is that it is relatively low-throughput compared to may of the other comparative techniques discussed here. Additionally, solved 3-D structures of the exact subject protein or a close relative to the subject protein are typically required.

Intrahost Variation Analysis

Next-generation sequencing technology has allowed for inexpensive deep sequencing analysis of individual viral infections. Due to high mutation rates, especially in RNA viruses, infection of an individual host will typically involve a

swarm of evolutionarily related viruses (55). The differences between intrahost variants are typically small, but with deep sequencing approaches they are identifiable (55, 56). This type of analysis can give researchers insights into the mutations related to host adaptation and immune evasion, among others. In one study of an RSV-infected child with combined severe immune deficiency syndrome, samples were taken before and after a bone marrow transplant to establish adaptive immunity (57). It was demonstrated through deep sequencing analysis that post-transplantation immune pressures began acting on the intrahost viral populations, revealing mutations within the immune-facing surface glycoprotein (the G protein) (57). This is just one example of the power of this type of analysis.

Phenotype and Metadata Analysis

As large-scale viral genomic sequencing efforts have become commonplace, the potential for detailed metadata and phenotype analysis has become possible. The practice of integrating detailed phenotype data with the genotype analyses is in its infancy. Although antigenic phenotyping of influenza virus is performed on a regular basis, this data is not routinely integrated into phylogenetic or cluster-based comparisons of viral data sets. Integration of metadata analysis into phylogenetic analysis is perhaps a bit more advanced. These types of analyses would be best used to determine if metadata could be associated with genotypes in any way, and perhaps to establish new phenotypes for viral species. Tools such as the Bayesian Tip-association Significance Testing (BaTS) exist to determine if specific metadata values have relationships to parts of or entire phylogenetic reconstructions (58). Similarly, the Geneological Sorting Index (GSI), as implemented on the molecularevolution.org

website, is another tool that is capable of examining the degree to which certain labeled metadata values group together within a specific set of phylogenies (59). Outside of a phylogenetic context, the Metadata-driven comparative analysis tool for sequences (MetaCATS) (32), as implemented on the IRD and ViPR bioinformatics resource websites, is capable of analyzing multiple sequence alignments to determine if specific amino acid or nucleotide variants can be statistically associated with particular metadata values. In addition to these tools, custom analysis is always possible with the use of statistical packages such as R or Matlab. As mentioned above, viral data standards are of the upmost importance. A cross-study analysis of metadata sets is only possible if metadata standards are adhered to when designing and implementing viral cohort studies (30, 31).

Study One: DASH: A Novel Method for Influenza Virus Surveillance and Vaccine Candidate Evaluation

Overview

This study aims to provide a computational framework for evaluating influenza vaccine candidates and demonstrates a model for a sequence-first approach to influenza virus surveillance (Figure 5). The data workflow described in this study provides a methodology for determining the specific viruses, based on virus sequence, which would provide the most information via hemagglutination inhibition (HI) assay experimentation in a hierarchical fashion. Additionally, the vaccine candidate evaluation methodology provided in this study extends the abilities of conventional antigenic distancing (AD) by providing a statistically robust approach to the association of antigenic phenotypes with viral genetic information.

State of the Art

Viral surveillance and vaccine candidate selection for influenza A and B viruses is a complex internationally coordinated process. Networks of clinical collection centers around the world take samples from patients presenting in their facilities with influenza-like illness (ILI). These samples are tested using the HI assays to assess their similarity to viruses known to have circulated in the recent past. This HI assay testing process is quite time consuming. Some of the viruses are fully sequences, but a subset of the viruses have only their HA and NA genes sequenced. The viruses selected for complete genome sequencing either demonstrate an irregularity with respect to the HI testing or are randomly selected from the cohort of HI tested viruses to provide genetic information for further characterization. In addition to the sequencing and HI testing, the HI assay data is used in a computational process known as antigenic cartography or distancing (AD). With AD, antigenic maps of viruses can be used to visualize and rudimentarily quantify the differences

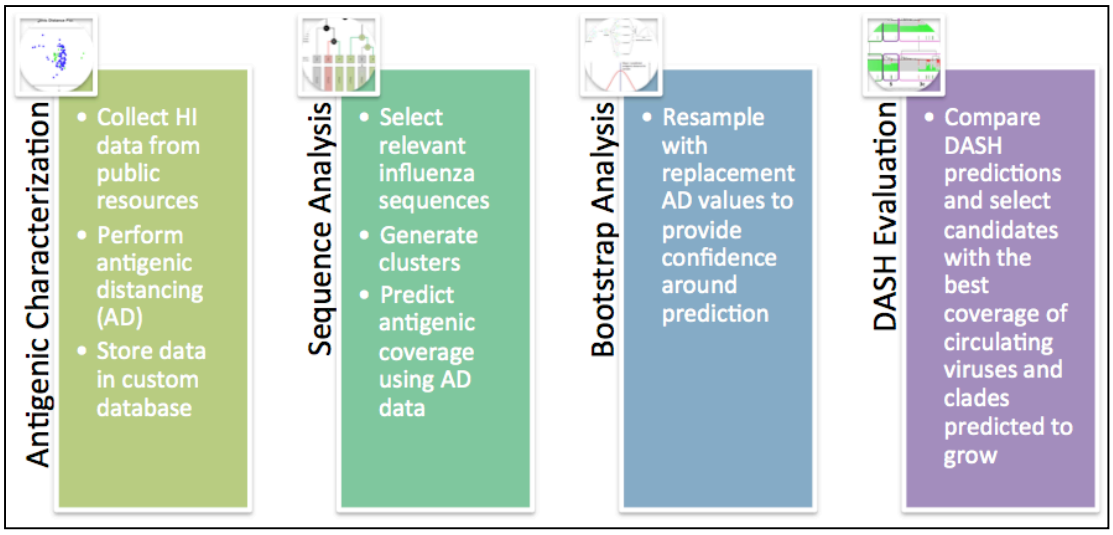


Figure 5 Computational Framework for Evaluating Influenza Vaccine Candidates.

between viruses. Direct associations between the antigenic phenotypes and other genetic information have, up to this point, been somewhat sparse.

This study hopes to extend the current state of influenza virus surveillance and vaccine selection by providing a framework for performing surveillance using a sequence-based approach first, followed by an HI assay approach. With this paradigm shift in surveillance methodologies we hope to extend the capabilities of current antigenic distancing technologies and reduce the latency between vaccine strain selection and the delivery of fully licensed vaccine products. Since the advent of NGS technologies, sequencing has become fast and cheap. All the genetic information from surveyed viruses can easily be captured using current approaches. From the genetic information, targeted HI testing plans can be developed using our methodology and can be used to enhance the antigenic information on currently circulating viruses. This study also aims to extend the capabilities of AD by providing a rigorous statistical assessment of the antigenic phenotype of all potential vaccine candidates. This statistical approach provides a more informed assessment of the antigenic relatedness of the viruses through a bootstrapping technique that accounts for the error in the AD methodologies.

Status of Work

The scope of work within this study has been incorporated into a manuscript that is currently in the revision stages. The findings of this study have been combined with the results from a novel complementary computational pipeline, known as proportion tracking. All analysis related to this study and proportion tracking have been completed and await consensus for publication.

Author Contributions

I developed the DASH pipeline, the grid-computing capabilities, early versions of uncovered cluster analysis R scripts and the antigenic coverage bootstrapping analysis. I also performed all of the initial protein clustering analysis; evaluation of the agglomerative hierarchical clustering techniques; designed, ran, and performed the retrospective analysis; and designed, ran and performed the computational component of the JCVI HI data analysis. I made minor suggestions for the improvement of the PTP pipeline, including use of the knee of the graph an appropriate height for cutting protein clusters. I developed the production pipeline wrapper around the preexisting PTP pipeline for expedited automated computation. I developed the name standardization methodology and the early database loading routines. I, along with other team members, manually curated the HI assay data. I, along with other team members, wrote and edited the relevant sections of the manuscript for publication.

<u>Study Two:</u> VirComp: A Novel Method for Viral Comparative Analysis Using Cluster-Based Gene Constellations

Overview

This study aims to enhance the offerings of the constellation or genotype analysis packages available for viral researchers (Figure 6). Viral genotype or constellation analysis is the systematic categorization of the entire genome content of one or more viruses based on predefined clustering criterion. The approach described here utilizes the farthest-neighbor clustering algorithm to build sequence dendrograms and clusters based on several predetermined percent identity cutoffs. This study also describes a novel near-publication ready figure generation application. Even though it has been designed for the visualization of viral genome constellations, this application could also be used for the visualization of metadata, phenotypes or other data types associated with viral or other biological projects.

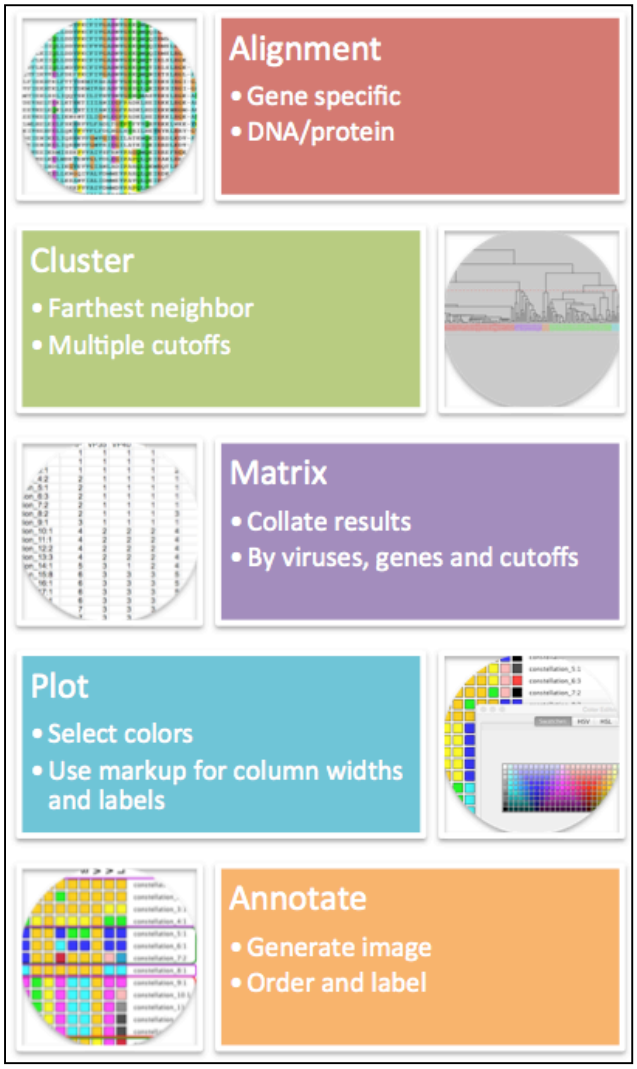


Figure 6 The Visualization of Viral Genome Constellations.

State of the Art

Virus specific packages and websites exist for performing constellation analysis, as well as the methodologies describing the process. Many of these applications rely on phylogenies to infer category membership and utilize phylogenetically derived criterion to cluster genes. Some of the available tools provide categorization within the context of a specific set of well-curated viral constellation typing databases.

It is the aim of this study to extend the available constellation analysis toolsets with a virus-agnostic methodology and a set of associated tools for clustering and visualizing constellations. This includes VirComp, a pipeline written in PERL that performs alignments, hierarchal clustering and cluster categorization of a set of viral genes, of a set of viral proteins, or of arbitrary subsequences within a genome of interest. These constellations are generated systematically in a virus-agnostic, context-specific manner using a sequence-identity based criterion. In addition to this novel methodology for identifying genome constellations, this study presents a novel tool for generating near-publication ready graphic visualizations of the constellation data using OrionPlot. In many cases the images could be used directly in a publication; although, the examples presented herein have been enhanced with relatively minor annotations to maximize impact and information gain.

Status of Work

The manuscript for this study has been completed and awaits final revisions before submission. In addition to the methods paper described above, specific

analyses using VirComp and OrionPlot have been published in two separate papers (4, 5). All code for VirComp and OrionPlot has been completed. The code has been deposited on GitHub and has been tested on a number of Mac OSX and Linux devices to ensure compatibility, to derive system requirements and to derive software requirements. All benchmark analyses have been completed.

Author Contributions

I, along with others, conceived of the VirComp constellation methodology. I wrote and refined the VirComp pipeline. I conceived of the OrionPlot program. I, along with others, performed the analysis of the swine, human, and avian influenza data sets. I performed the analysis of the RSV and Ebola data sets. I performed the comparative analysis with phylogentic techniques. I, along with others, wrote the manuscript.

<u>Study Three:</u> Large-Scale Respiratory Syncytial Virus Whole-Genome Sequencing Identifies Sequence Duplication in G Gene Associated with Reduced Diseases Severity

Overview

Study three aims to examine genetic determinants for the pathogenicity and epidemiology of respiratory syncytial virus (RSV) (Figure 7). Through an integration of phylogenetic methods and statistics, this study uncovers potential determinants for disease severity and transmission based on associations with clinical metadata. These

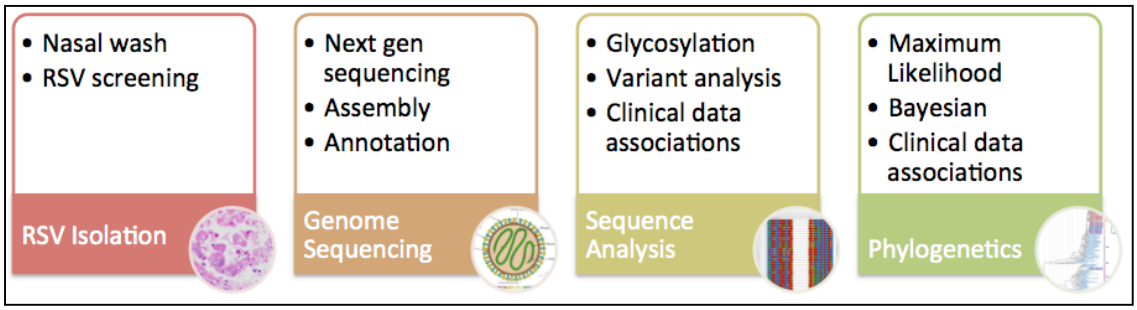


Figure 7 Genetic Determinants for the Pathogenicity and Epidemiology of Respiratory Syncytial Virus.

determinants range from the discovery of positive selected sites within several RSV genes, to glycosylation patterns on the G protein and examinations of large inserted repeat sequences.

State of the Art

RSV epidemiological surveillance is currently in its infancy. No globally-coordinated whole-genome surveillance efforts are currently being implemented. In the past few years, large numbers complete RSV genomes have become available. As of 2015, more than 600 genomes have been released to public data repositories. The deposition of large amounts of complete genomic information enables researchers, for the first time, to examine these data for molecular signatures of epidemiology and pathogenicity.

In an exploratory context, this study extends the current knowledge of RSV by integrating clinical metadata with phylogenetic analyses of whole RSV genomes, as

well as individual RSV genes. This study also aims to enrich the RSV research by making associations between the signatures of molecular evolution and the epidemiological or pathogenic phenotypes.

Status of Work

The work from this study will be broken into two publications. The first paper focuses on large-scale phylogenetic trends in RSV and the inferences that can be made between the genetic data and the clinical metadata. The second paper will focus on the specific selective and mutational patterns in the various RSV genes, as well as an analysis of the glycosylation patterns in seen in the study data set. The first paper has been drafted and is in the revision stage. The second paper has been conceived of and exists in text only as Chapter 4 of this dissertation. All analyses for these two papers have been completed.

Author Contributions

I, along with others, conceived of this study. I, along with others, designed the phylogenetic experiments. I performed neighbor joining, maximum likelihood and Bayesian analyses. I performed selection and variation analysis. I performed the glycosylation analysis. I performed the clinical data statistical analysis and BaTS analysis. I, along with others, wrote and edited the publication manuscript.

Chapter 2: DASH: A Novel Method for Influenza Virus Surveillance and Vaccine Candidate Evaluation

Abstract

The influenza virus poses a significant risk to human health. Up to 500,000 people are estimated to die annually during the influenza epidemics of the Northern and Southern Hemispheres. The World Health Organization (WHO), and its related network of surveillance facilities, seasonally monitor the evolutionary and antigenic signatures of the viruses circulating in human populations around the globe. This viral surveillance informs the semi-annual recommendations that are made to the public and pharmaceutical industry on the formulation of the seasonal influenza vaccine. The data for this process takes months to collect, and it can take up to nine months for the pharmaceutical industry to turn the recommendations into a vaccine product ready for dissemination to the world population. Here we have described a method, Distancing of Antigenicity by Sequence-based Hierarchical Clustering (DASH), which may alleviate some of the data analysis burden on those involved in viral surveillance. DASH is capable of making statistically grounded choices for vaccine candidates based on a phenotype mapping algorithm and protein clustering. Each vaccine candidate can be evaluated for its suitability as a component of the influenza vaccine and compared against other antigens. Similarly, through a close monitoring of available antigenic data derived from the hemagglutination inhibition (HI) assay, DASH can produce lists of viruses that would either enhance our surveillance of circulating influenza viruses or be used in a manufacturing at-risk

context for the pharmaceutical industry. Coupled with reverse genetics, cell grown vaccine components, and synthetic biology, it is our belief that the DASH methodology will enhance preparedness for seasonal influenza epidemics and shorten the manufacturing latency between strain selection and vaccine administration.

Introduction

Influenza is a significant risk to human health. It is estimated that hundreds of millions of people are infected annually, causing between 250,000 and 500,000 deaths (60). The virus is a negative-sense single-stranded RNA virus with eight genomic segments that code for 10-12 known proteins (61). Of these proteins, the hemagglutinin (HA) and neuraminidase (NA) proteins are of most interest immunologically, as these proteins are on the surface of the virus and are directly accessible to the host immune system (61, 62). Influenza A viruses (IAV) are currently categorized into 17 HA types and 9 NA types (61). Combinations of the two proteins result in the virus subtypes that are commonly known, such as H3N2, H1N1 and H5N1. Influenza in humans is primarily caused by two types of influenza viruses – influenza A virus (H3N2 and H1N1 subtypes) and the influenza B virus (IBV) (60). In the case of IAV there is the potential for zoonotic transfer from an avian host reservoir to humans (63), as illustrated by H5N1 (64). All human influenza viruses (H3N2, H1N1 and influenza B) are under constant selective pressure to evade the host immune response and therefore undergo constant antigenic (and genetic) drift (65) to form new variants – some with epidemic, and occasionally pandemic potential (60). At the present time, the best way to combat influenza infections is by prevention through an extensive worldwide vaccination program (60,

64). For these reasons, influenza surveillance must be two pronged (64). First, to track the continual drift in the currently circulating strains, robust genome sequencing and antigenic tracking surveillance must be performed (64). Secondly, to properly prepare for a pandemic of a previously zoonotic origin virus, such as H5N1, sequence and antigenic surveillance must also be performed in animal species known to carry and transmit influenza (66).

With these surveillance data, it is possible to extensively analyze and, in some cases, build predictive models that aid in the production of multivalent influenza vaccines (42, 67-69). Current methods for selecting vaccine candidate viruses have focused on a combination of techniques. Phylogenetic analysis and tracking of co-circulating viruses of different genetic lineages using the sequence surveillance data for the HA gene (and sometimes the NA gene) is part of the process (60). Additionally, antigenic distance (AD) mapping (42) (sometimes referred to as antigenic cartography) of select strains using the hemagglutination inhibition (HI) assay is also a major component of the selection process. The latter does not rely on sequence data; however, for proper correlation of genetic lineages, or clades, to a particular antigenic group, some overlap in sequence and HI data sets is required (42). Once it is determined which virus clade is dominant and the antigenic characteristics of that clade during any given season, a recommendation to change or maintain the current viral components of the multivalent vaccine can be made. This decision is further informed by the phenotypes of the virus related to growth and suitability for production in eggs (60, 70). Although the general framework for this semi-annual decision is understood, the fine details of this process are not publically available. By

this method, it can often take months to properly calibrate the reagents being used in the production process. This method can prove to be too ineffectual during the course of a pandemic to prevent the spread of infection, as was seen in the H1N1 pandemic of 2009 (64).

Bioinformatic techniques have previously have been used to study the evolution and antigenic traits of influenza. These techniques have begun to make the selection of vaccine candidates easier; however, the current speed at which vaccine candidate reagents can be manufactured and calibrated is a limiting step in both pandemic responses and seasonal vaccine production (64). Here we demonstrate modifications to the current methods that aim to speed the vaccine production process and allow better understanding of the nature of antigenic drift. By tuning bioinformatic approaches to look for potential candidate viruses for manufacturing at-risk (viruses that show some evidence for utility as a vaccine candidate, but not fully validated yet) and by further qualifying those candidates using computational inferences, we can greatly speed the selection of vaccine candidates for production. Coupled with synthetic biology and reverse genetics technologies (not discussed here), new vaccine reagents could be produced in days to weeks, rather than months (64).

Methods

Data Collection and Database

We have created a database of all publically available human IAV H3N2, H1N1 HA sequences (seasonal and pandemic), and IBV HA sequences found at both the GenBank and EpiFlu (Figure 8). The database is a relational database implemented using MySQL (version 5.6.25 available from www.percona.com) relational database management system (RDBMS). The HA sequences have been processed such that

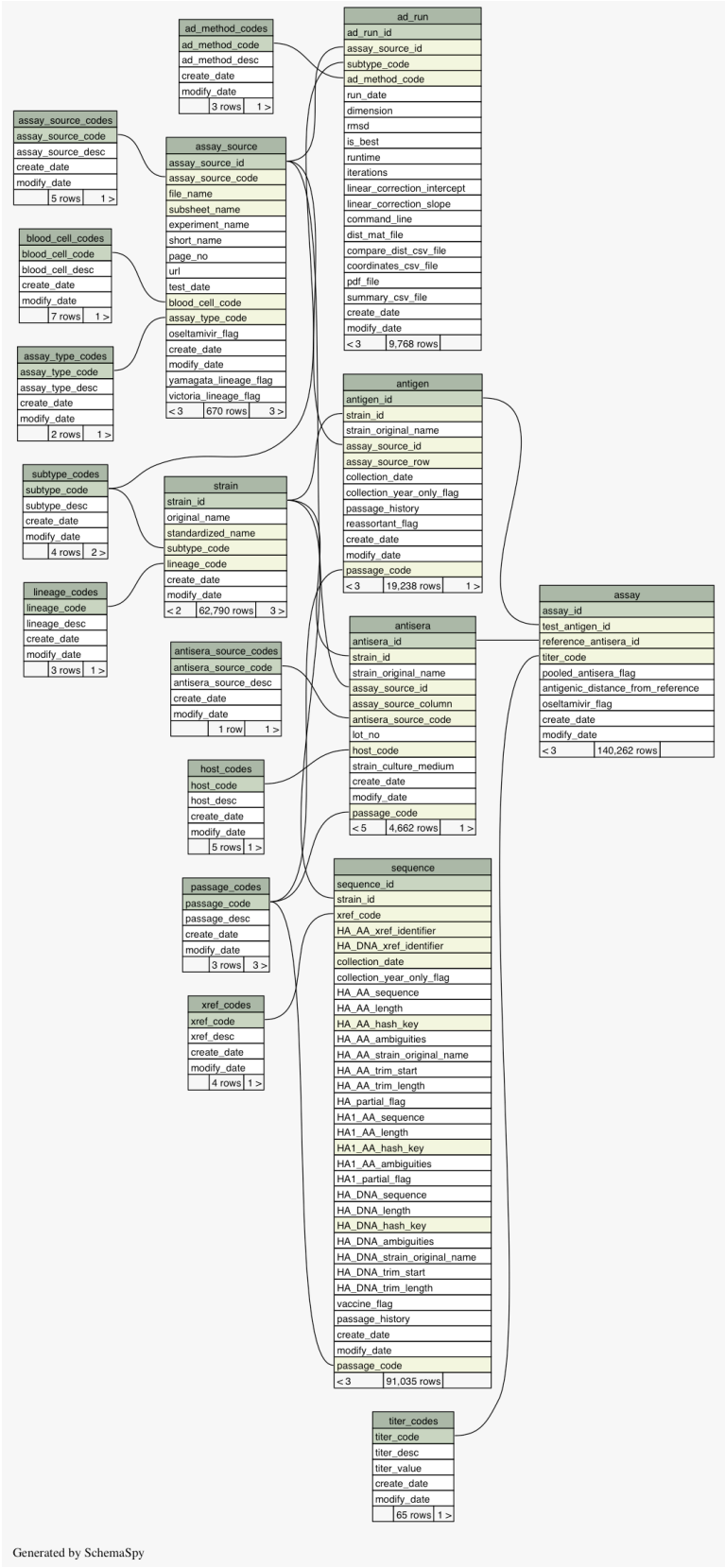


Figure 8 Schema of Custom Influenza Sequence and HI Data Database. The central entity in this schema is the strain table representing a viral strain. Each strain is linked to its associated sequence and HI data via foreign key relationships. It is essential to properly recognize synonymous strain names to preform this link.

protein and nucleotide sequences from these databases have been paired together. We further categorized these sequences according to their completeness, full-length coding DNA sequence (CDS), full-length HA1 domain, or partial sequence. Metadata was collected on all sequences, as well as the passage number, passage type (cell or egg), and collection date. Strain information and passage information were standardized to allow for more accurate matching of strain sequence and HI data. Standardization was performed on geographic locations, lab identifiers and collection years. For example, A/CA/07/09 would be standardized to A/CALIFORNIA/7/2009. It was our practice to use full geographic locations wherever possible, to remove leading zeros from lab identifiers, and to expand two digit years to full four digit years. This ensures that a sequence labeled A/California/7/09 would accurately be associated with HI data labeled A/CA/07/2009, since both standardized names would resolve to A/CALIFORNIA/7/2009. Geographic standardizations were done using the Google Maps application program interface (API). Ambiguous results were curated manually. The database stored both the standardized and the original names to allow us to track possible errors in associations, if discovered. Similarly, passage metadata was mapped to egg or cell passage type via a mapping file that was manually curated. For example, SIAT1 was mapped to cell passage type, since SIAT is a cell culture line. Conversely, E1E5 was mapped to egg passage type, since E represents an egg passage. If a passage was ambiguous, such as SIAT1E5 (SIAT1 for cell and E5 for egg passages), we mapped it to egg – as even a single egg passage has been shown to be enough to shift a human virus toward avian adaptation, potentially altering its antigenicity.

In addition to the sequence data we populated our database with an extensive set of HI assays. HI data was collected from both the CDC via annual Vaccines and Related Biological Products Advisory Committee (VRBPAC) reports to the Food and Drug Administration (FDA), and semi annual National Institute for Medical Research (NIMR) (in the UK) surveillance reports to the WHO. These HI data tended to be highly variable in format, with numerous errors due to the likely manual method by which they were created for their respective reports. We therefore manually reviewed these tables for consistency of format and content to minimize the errors loading into our database. Once loaded, we stored the metadata on the HI run (such as the source and test date), as well as the metadata on the strains (such as passage and collection date information). A combination of passage information, collection date, and strain names were used to associate specific HI assay distances with specific sequences obtained from GenBank and EpiFlu. This step was critical, as passage differences often alter sequences and affect the HI assay distances for a particular virus.

The final component of the database was the storage of AD data results from our antigenic distancing pipeline. The AD pipeline will be described later; however, the results included the AD distance matrix data for 15 separate AD runs in various dimensional spaces, using various distance calculation methods.

Access to the database was provided via a PERL API and command line scripts, and later incorporated directly into the DASH pipeline to ensure seamless automated computation. PERL scripts were also used for the database loading process. The current database includes 62790 strains, 57609 complete HA sequences, 62914 HA1 sequences, and 33426 partial HA protein sequences. Broken down by

subtype, there are 36778 H3N2 viruses, 8657 H1N1 viruses, 29484 H1N1pdm viruses, 7764 IBV Yamagata, and 8352 IBV Victoria sequences in the database collected between 1918 and January 12, 2015. Additionally, there are 670 HI assays for a total of 140262 pairwise antigenic distances.

Antigenic Distancing

We used a technique known as antigenic cartography, herein described as antigenic distancing (AD).

Hemagglutination Inhibition Assay – Antigenic distance is measured using the HI assay. In this assay, a solution containing red blood cells is combined with an influenza viral isolate and antiserum made in a mammal using a different virus. First, the antisera are serially diluted and then a fixed amount of virus and blood cells are added to each sample (42). If the antibodies in the sera do not bind the virus, the blood cells will bind the influenza virions and form a lattice, i.e. hemagglutinate. However, if the antibodies in the serum bind the virus, no hemagglutination takes place. This inhibition of hemagglutination is dependent on both the intensity of the virus-antibody binding and the antibody concentration. This test is repeated with the opposite combination of virus and antiserum. These data are compared to determine the viral titer required to prevent red blood cell agglutination in the presence of antiserum. Antigenically similar viruses will fail to agglutinate red blood cells at a similar titer (within 4-fold) (71) when in the presence of their inhibition partners' antisera. In addition to two-way titer assays, one-way assays are also performed to survey circulating strains to identify low antibody-antigen viral reactors that may indicate antigenic drift.

Homo and Heterologous Distance Calculation – Titer data from the above-described HI assay were converted into antigenic units using the following two formulas.

(1) Two-way $d_2 = \sqrt{(H_{11} H_{22}/H_{21} H_{12})}$

(2) One-way $d_1 = \log_2(H_{11}/H_{21})$ (71)

Antigenic Distancing via Multidimensional Scaling – Through antigenic cartography HI data can be plotted in 2-dimensional space to show the relative antigenic distances between multiple influenza strains. Once the above distances were calculated from HI assays, the antigens were plotted in a multi-dimensional space and the two-way pairwise distances between all members were calculated using a process known as multi-dimensional scaling (MDS) (42). We believe that 2-dimensional distancing is an oversimplification of the data for the sake of visualization, and thus have used a modified version of antigenic cartography to calculate antigenic distance. We have utilized several different distance methods to calculate pairwise distance in multi-dimensional space rather than the simple 2-dimensional calculation used in traditional antigenic cartography (42). We have chosen to employ multi-dimensional antigenic distancing because 2-dimensional calculations do not take into account the multi-dimensionality of the HI data. With each new distinct antiserum that is introduced into an HI assay an additional dimension is added to the data. There are several methods for calculating the distance between two points in a multi-dimensional space. We utilized three methods: Euclidean, Manhattan, and Minkowski, and evaluated each for its overall effectiveness. The dimensionality and distance method that provided the best fit, as

measured by root-mean squared deviation (RMSD), for a particular set of HI data was selected by DASH for the next step in our analysis procedure (Figure 9).

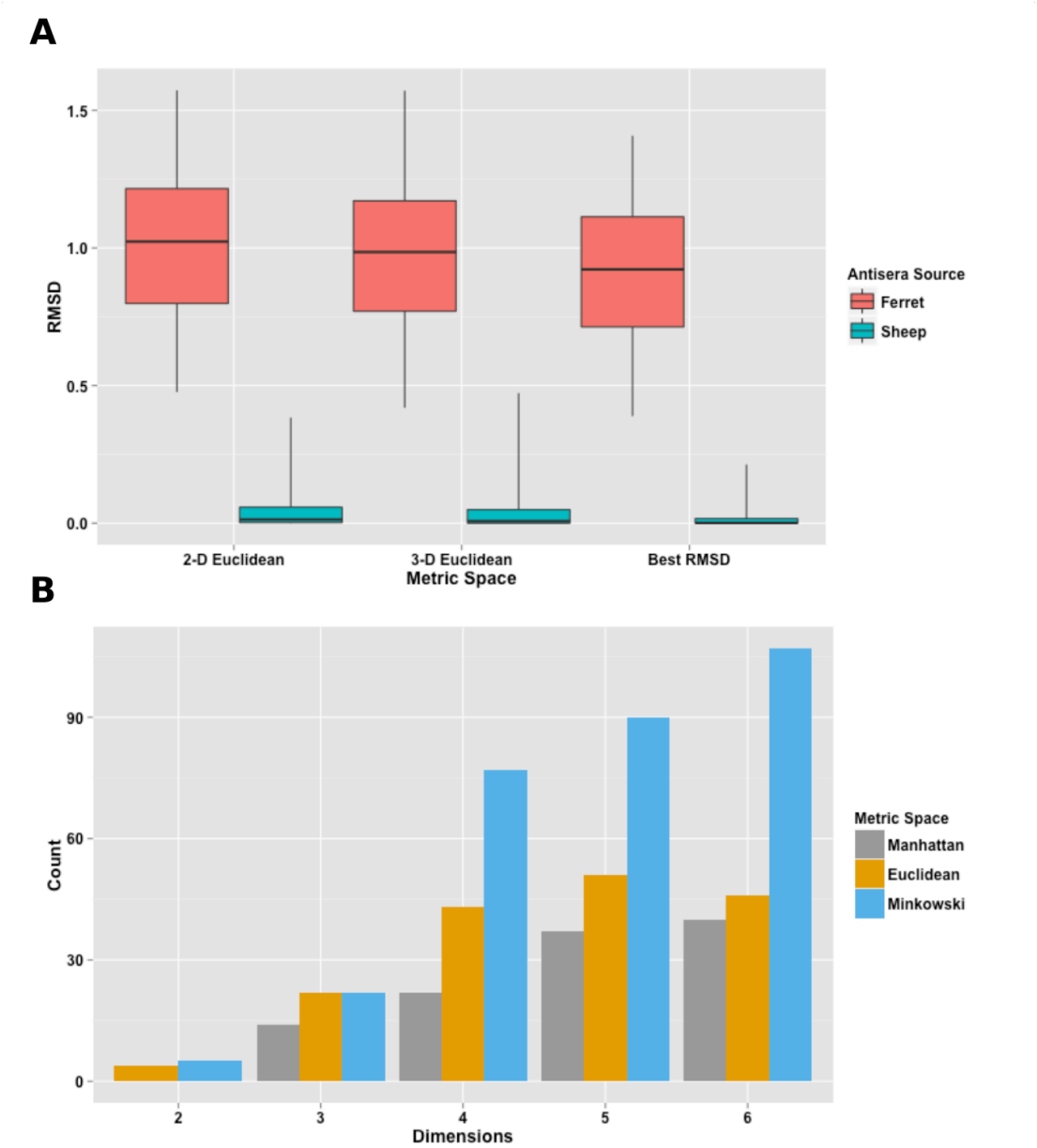


Figure 9 Comparison of RMSD Values for Various Antigenic Distancing Experiment Using Mixed Metric Space and Dimensionalities and Frequency of Use. (**A**) Box and whiskers plot showing the 25-75% interquartile range (box) and 2.5-97.5% range (whiskers) of the 580 AD runs of ferret antisera and 164 AD runs of sheep antisera HI assay data. Statistically significant differences between the ferret 2-D to Best RMSD and 3-D to Best RMSD exist (p-values = 1.99e-09 and 4.761e-05 respectively) indicate our method of selecting the best of 15 different distance metrics and dimensionalities produces better results than traditional 2-D or 3-D Euclidean methods. (**B**) Histogram plot of ferret sera Best RMSD metric and dimensionality over 580 HI assay data sets. This result indicates higher dimensionality and often the Minkoswki metric space as a more accurate measure of antigenic distance via the RMSD of the AD plot.

Distancing of Antigenicity by Sequence-based Hierarchical Clustering (DASH)

Pipeline architecture overview **–** The DASH computational pipeline consists of a series of PERL and R scripts (Figure 10). Data is prepared and permuted using the various scripts. Then, together with the results of various analyses, the data is compiled for inspection by analysts familiar with the methodology. The following will describe the pipeline architecture, based on three phases of computation.

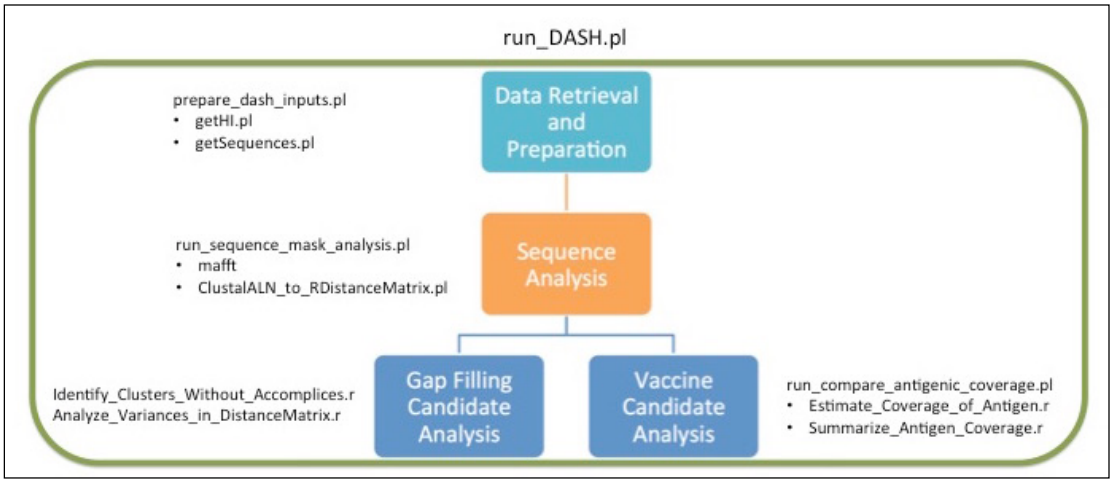


Figure 10 Workflow of the DASH Pipeline. Scripts responsible for various aspects of the computational pipeline are indicated next to category boxes in the diagram.

Data extraction and sanitation **–** The first step of the run_DASH.pl script is to extract all the relevant information needed to perform the DASH analysis. The user provides the date ranges for which they wish to analyze data and run_DASH.pl will in turn run getHI.pl and getSequence.pl to extract the relevant sequence and AD data from the MySQL database. Data sanitation is required to ensure AD data and sequence data is in a format necessary to link the data types together. Specifically, egg and cell passage labeled sequence and AD data are properly tagged with this information, such that linking of the sequence and the AD data can occur during later

analysis steps. If the passage data is ambiguous for either the AD or the sequence information, the passage information is stripped from the record. The run_DASH.pl script also automatically creates lists of antigens to run vaccine coverage predictions on. The criteria for inclusion in this analysis set is that there exist at least 40 pairwise AD across the subset of HI data included in the analysis, thus providing a ≥ 95% prediction interval for its antigenic distances with respect to other strains in the same season. We have specified 40 distances, as our cutoff, since smaller subsets of AD data tends to not give accurate estimations of predicted coverage because there are not enough viruses with both sequence and HI data available to anchor the phenotypic coverage predictions (Figure 11). This cutoff requirement has not been rigorously evaluated; however, larger protein cluster dendrograms tend to require more HI AD data to make plausible inferences across the entirety of the clusters. Similarly, unevenly tested antigens (antigens that have been tested heavily against viruses from one cluster, but with little or no testing against viruses from a different cluster) show more variability in the coverage prediction results.

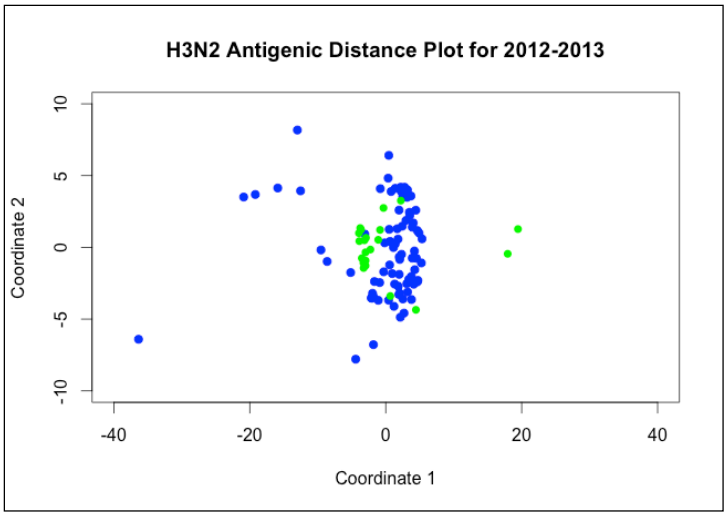


Figure 11 Antigenic Distance Plot of Three HI Assays Performed for The 2012-2013 Vaccine Strain Selection. It should be noted that to generate representative distances during a given season, multiple HI assays should be performed. The data points in green, for instance, do not necessarily represent the entire population of distances possible during the 2012-2013 season.

ANDES scripts – After data extraction and sanitation is complete, run_DASH.pl invokes the main DASH script (run_sequence_mask_analysis.pl). This script is responsible for running the majority of the preliminary analysis R and PERL scripts. The run_sequence_mask_analysis.pl script also runs the run_estimate_vaccine_coverage.pl script. The run_estimate_vaccine_coverage.pl is responsible for running the individual coverage predictions for the set of antigens with sufficient pairwise HI AD data, as established by run_DASH.pl. Each prediction takes several minutes and often there are tens to several hundred antigens to analyze; therefore, to speed up computation we have enhanced DASH with grid computation capabilities.

GRID computing enhancements – The run_estimate_vaccine_coverage.pl is grid enabled and launches individual vaccine candidate coverage predictions, as well as the bootstrap analysis on the JCVI grid, supported by the SunGrid Engine. The summarization of antigenic coverage predictions is dependent on all the grid computations being complete, thus run_estimate_vaccine_coverage.pl has grid-monitoring capabilities built in. Once summarization of all predictions is complete, DASH analysis has finished.

Pipeline Details

Alignment – An important step in the DASH computational pipeline is to create protein clusters. To do this, proteins are aligned and the pairwise distances are calculated for them using the program MAFFT (set with default options) to create the multiple sequence alignments (72, 73). MAFFT is a fast method for performing

sequence alignment. With such highly similar sequences, as are seen in the various influenza types we have analyzed, the probability of misalignments is low.

Modified BLOSUM62 – DASH uses BLOSUM62 to calculate distance between the subject protein sequences. Although, technically not a distance matrix, we have derived a distance matrix from the BLOSUM62 scoring-matrix. To make this derivation, similarity scores were rescaled between 0 and 1 and inverted to represent distances.

Protein Masking – For the same reason as the choice to examine antigencity on protein clusters, as opposed to phylogenetic clades, it may be possible to find portions of the HA protein that are critical in determining the antigenic phenotype of the virus (62, 65, 67, 69). DASH uses the concept of protein masking to eliminate from consideration certain non-relevant areas of the HA protein when calculating distance. DASH implements protein masking in the following ways to ensure just the

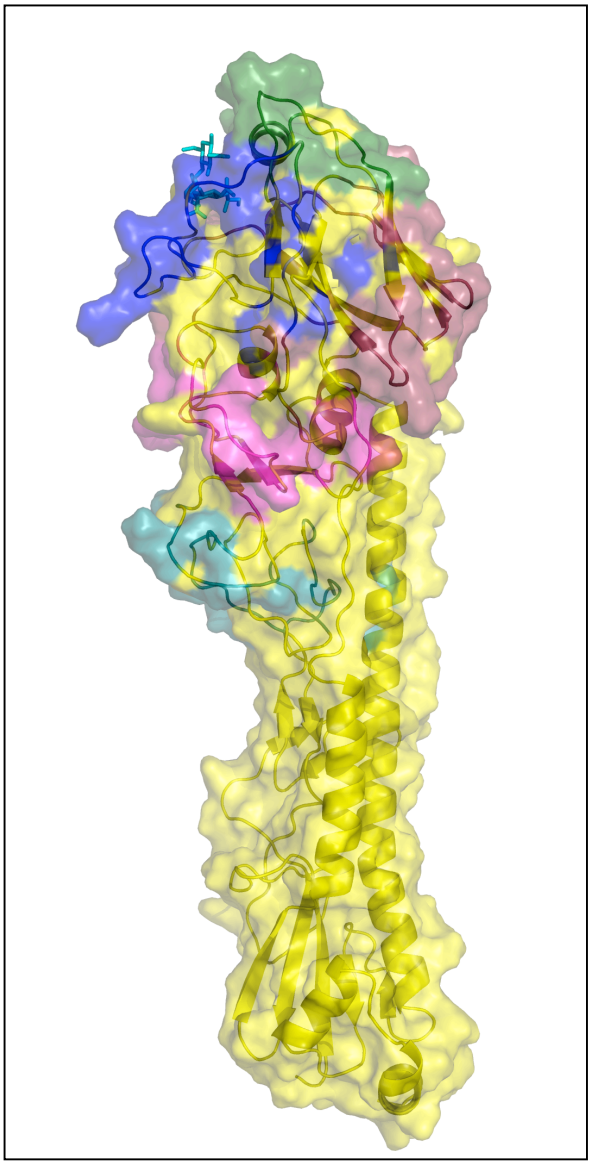


Figure 12 Influenza H3N2 3-D Protein Structure With Five Known Immunodominant B-Cell Epitopes Highlighted. The known epitopes are colored according to site A, B, C, D, and E (green, blue, maroon, pink, and teal respectively).

relevant specified sequence is considered for the distance calculation: all HA, HA1 domain, all known antibody epitopes, and specific known antibody epitopes (H3N2 AB epitopes and H1N1 SaSb epitopes) (Figure 12). The masks were developed through the curation of HA protein data using the literature citations that documented the antibody binding sites of the various influenza types.

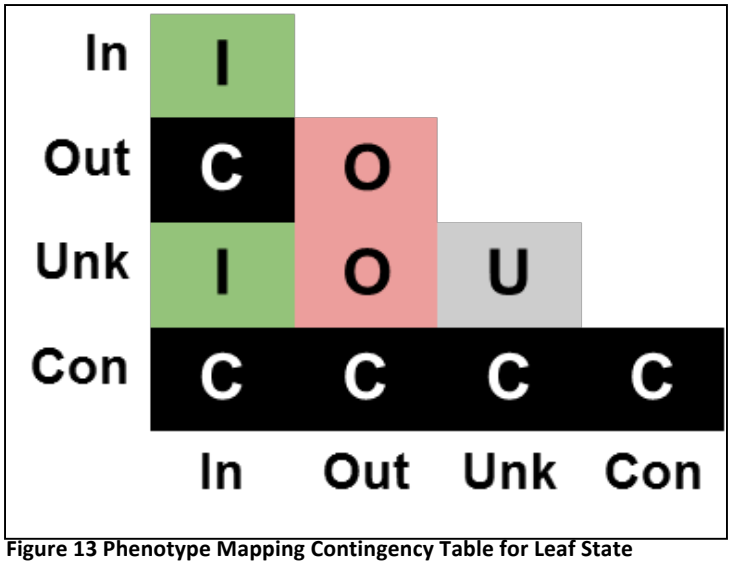


Figure 13 Phenotype Mapping Contingency Table for Leaf State Inheritance

Hierarchical Protein Clustering – Clustering HA protein sequences is an essential part of the methods described here. DASH uses an agglomerative hierarchical clustering method that is deterministic. This allowed for repeated clustering experiments on the same data with the same resulting dendrograms. Specifically, DASH utilizes the Ward's minimum variance method (74) as implemented in the R package of statistical software to cluster the protein sequence data.

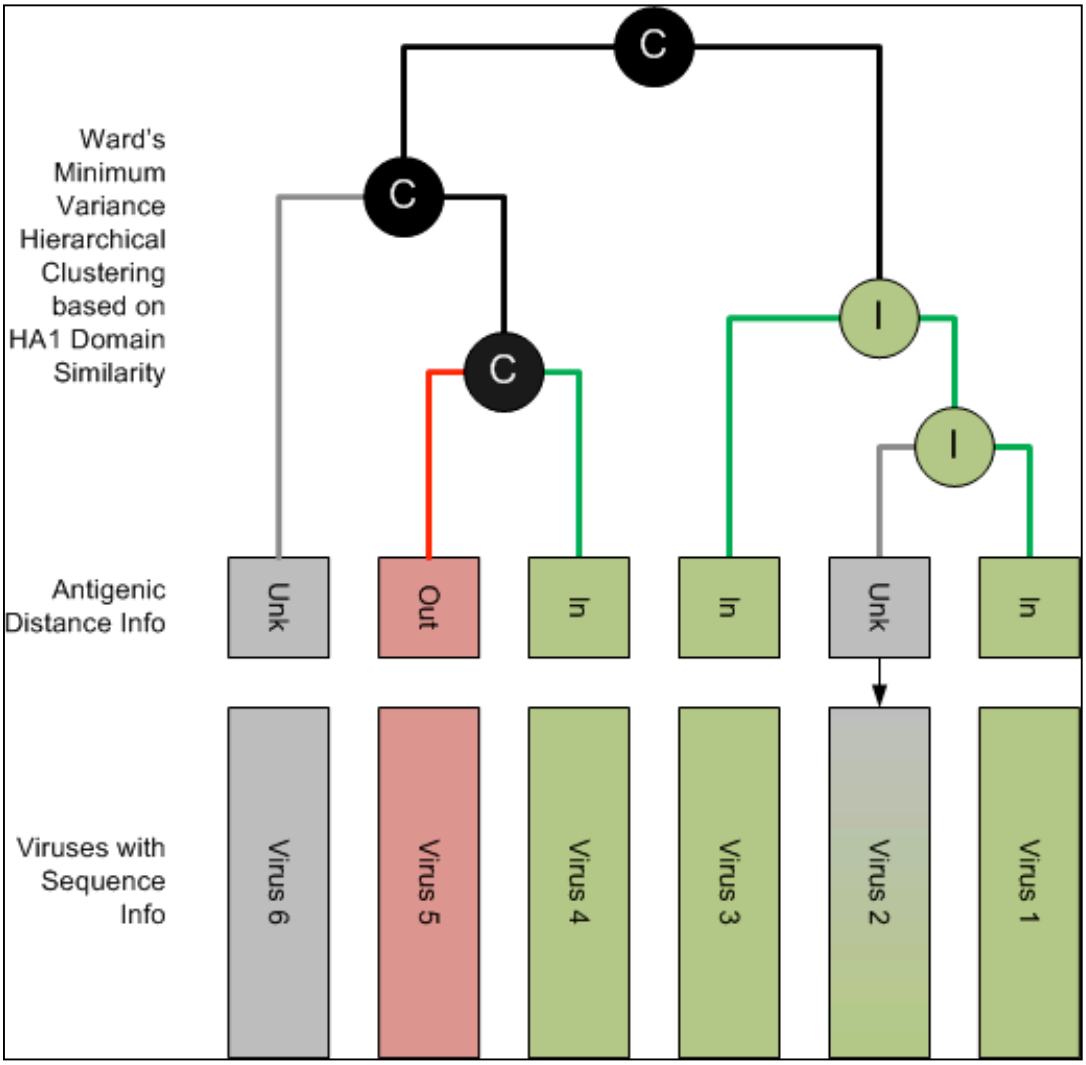


Figure 14 Inference Algorithm Utilized by DASH. Each node in the dendrogram is assigned a state (In, Out, or Conflicted) based on whether or not its decedents contain leaves with HI data points and whether those leaves conflict. If no conflicts exist all decedents from that node will inherit the parent nodes state. If conflicts exist an unknown state will be assigned to all non-HI leaves descending from the common node marked as conflicted.

Phenotype Mapping – To predict the effectiveness of particular vaccine seed

candidates, DASH uses a novel technique of mapping antigenic phenotype traits onto the protein clusters generated using the methods described above. The mapping process focuses on one single viral antigen at a time and multiple candidates are evaluated in parallel using grid-computing technology. Distances generating by the AD process are used as well as the protein cluster data. If the reference antigen is determined to be antigenically similar to the candidate antigen it will be assigned to an "in" group. If the reference antigen is determined to be antigenically variant it will be assigned to an "out" group. Antigenic similarity, or "likeness", is most commonly estimated to be ≤ 2 AD units. DASH projects this "in"/"out" grouping onto the protein cluster dendrogram, and the leaf nodes will be colored according to a prediction associated with the "in" or "out" group (Figures 13 and 14). For a leaf node to be colored it must be fit several criteria. First, if it is an HI accomplice

sequence, meaning we have both sequence data and HI data for the leaf sequence, it will be colored according to the group to which it belongs. Next, if a leaf node is between two accomplice leaves of the same group and there are no leaves of a conflicting group at the same level in the hierarchical cluster, it will be colored with the same group as the two accomplice nodes. Lastly, if a terminal cluster (a cluster extended far off the trunk of the dendrogram with no additional branches) contains one accomplice and there are no conflicting nodes at the same level outside the terminal group, all nodes in the cluster will be colored with the group of the single accomplice in the cluster. These colored predicted antigenic coverage protein clusters form the basis of DASHs vaccine candidate selection algorithm described below.

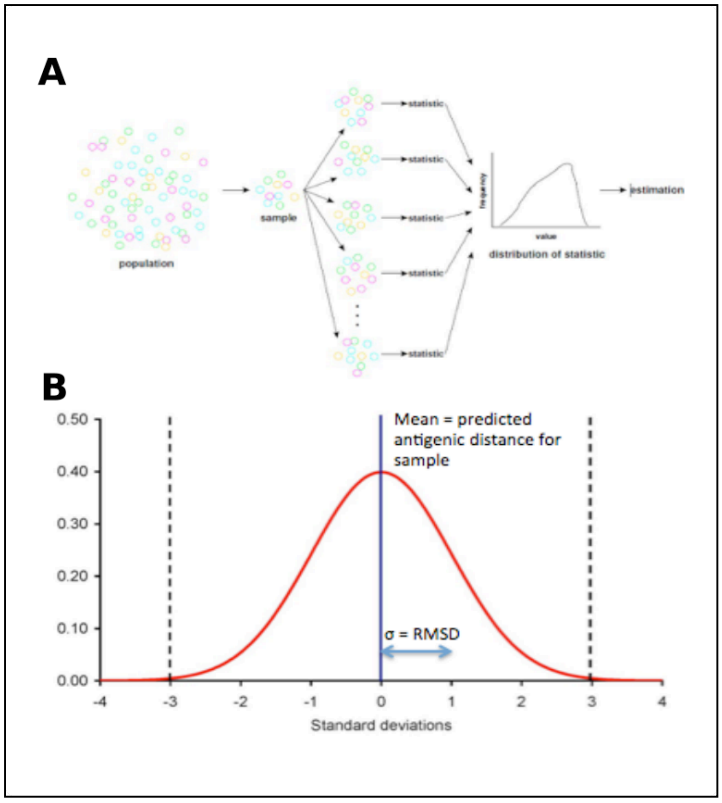


Figure 15 Bootstrap Analysis Methodology Used to Assess Reliability of Antigenic Coverage Predictions. (**A**) Bootstrap analysis uses random resampling with replacement from a sample population. In the case of DASH the resampling is done on the set of Antigenic Distances (AD) available for a particular influenza season. (**B**) Once the HI data point is sampled it is modified by the DASH algorithm. DASH selects a new value for this data point randomly from a distribution with a mean equal to the original AD value and the standard deviation equal to the RMSD of the MDS computation where it was generated. This RMSD approximates error associated with performing both HI assays and the AD algorithm.

Bootstrap analysis – It is important to establish the statistical accuracy of the predicted vaccine seed candidate. To do this, the phenotype-mapping

algorithm includes a bootstrap analysis to measure the variance of the predictions. Bootstrapping is a process by which a population sample is resampled with replacement repeatedly (Figure 15A). With each resampling, statistical measures are made (such as median or mean). After several iterations confidence intervals around the original statistic can be generated from the bootstrap replicates. DASH employs a modification of this technique. After an initial phenotype mapping step, the population of pairwise antigenic distances is resampled with replacement. Instead of taking the original distance value; however, DASH resamples the distance over a normal distribution where the mean is defined as the original distance and the standard deviation is the RMSD value of the AD computation used to generate that distance originally (Figure 15B). It is this modified resampling that allows DASH to take into account the error associated with both the HI assay and the AD computation when making coverage predictions.

Identification of Clusters without AD data – During the initial phase of DASH analysis, identify_clusters_without_accomplices.r identifies 10 clusters without protein sequence data but with AD data. These clusters are of the utmost importance to surveillance because they represent newly emerging genetic variants. Antigenic testing on these newly emerging clusters is key to maintaining high levels of phenotypic measurements across the entire diversity of circulating viruses. These new uncovered protein clusters are then analyzed using the R script (Analyze_Variances_in_DistanceMatrix.r) to established the HA sequences at the center of the cluster (the centroid virus), or those viruses closest to the center of mass of the sequences being examined in the uncovered cluster. We have dubbed these

centroid viruses DASH surveillance candidates, as further antigenic testing should be performed on these viruses for maximum gain of phenotypic information. In a later experiment, DASH surveillance candidate viruses were constructed at JCVI with synthetic genomics technology and tested in HI assays, as a proof-of-concept that this methodology could be used to drive influenza virus surveillance in real time and positively influence and refine DASH predictions. In certain instances, these centroid viruses show the antigenic characteristics of good vaccine candidates. Retrospective analysis suggests that without running HI analysis on the centroid viruses, selecting these DASH surveillance candidates from several uncovered clusters often leads to the prediction of WHO vaccine or manufacturing candidates prior to the release vaccine recommendations.

Identification of Vaccine Candidates – DASH uses grid-computing technology to generate candidate antigen centered phenotype maps. Each antigen that has undergone antigenic distancing analysis and meets the DASH minimum distance requirement is evaluated as a potential vaccine seed in the phenotype mapping analysis. A bootstrap analysis is also performed for each mapping run to further establish the predicted antigenic coverage of that antigen against similar, contemporary viruses. An effective vaccine prediction maximizes the predicted antigenic coverage while simultaneously minimizing the variance in the prediction. During the retrospective and JCVI HI analysis experiments statistical analysis of all qualifying antigens and their bootstrap analyses were performed to establish the candidate seed viruses that fulfill these criteria.

Proportion Tracking Pipeline

In parallel to the development of DASH we developed a methodology to track the proportions of the subpopulations of influenza viruses circulating concurrently during any arbitrary time period. We call this method the Proportion Tracking Pipeline (PTP). These subpopulations are not inferred using phylogenetic methods and do not necessarily follow the standard clade designations provided by the CDC and the WHO. Instead the subpopulation designations are generated using the Ward's clustering method, as it is used in DASH. This allows us a common clustering methodology between DASH and PTP that provides consistency when comparing PTP results to DASH predictions. Knowing which groups of viruses are expanding or contracting allows us to put the DASH analysis into context and to determine which DASH predictions should be emphasized. To perform the PTP analysis, first all HA1 domain sequences from a particular subtype or lineage of influenza viruses are partitioned into 18-month subsets of viral sequences. Each partition overlaps by 12 months in a windowed fashion based on the collection date for the virus. This results in 12 months of viruses seen in other partitions and 6 months of novel viruses. Next, we apply weighting to the viruses in each 18-month partition. Viruses isolated in the oldest 6-month period are given a weight of 0.16, viruses from the middle 6-month period are given a weight of 0.33, and viruses from the most recent 6-month period are given a weight of 0.50. Clustering is performed and a cutoff for the partition is calculated. This cutoff is derived by calculating the sum of squares for the intra-cluster group distances at each possible height (k) and dividing it by the sum of squares for the distances for entire set of sequences. Each

value is plotted; and the k value nearest the knee (or inflection point) of the resultant graph is selected and used to generate clusters for the partition of interest. This process is repeated for all partitions over the time period being examined. Once clustering is competed, the proportions for each cluster are calculated using the weighting scheme described above. Clusters between partition periods are linked based on similarity, to show forward associations of past partitions' clusters. Both the weighting scheme and the linkages between partitions provide continuity in the analysis, showing clusters that continue to expand over successive partition periods versus those that shrink or die out altogether. Clusters with the most of expansion during a given period are weighted accordingly, allowing for visualization of the importance of a particular viral subpopulation.

Retrospective Analysis

We performed a retrospective analysis over 11 years and 22 influenza seasons for IAV H3N2, IBV Yamagata, and IBV Victoria. For the IAVs H1N1 and H1N1pdm, we analyzed seven years (15 seasons) and four years (seven seasons), respectively. The retrospective analysis involved pulling sequence and precomputed AD data from a 12-month period for each influenza season examined. For the Northern Hemisphere seasons, the dates ran roughly from March to March. For the Southern Hemisphere seasons, the dates ran from September to September. The month cutoffs for analysis corresponded with the WHO's Northern and Southern Hemisphere vaccine selection meetings. We are assuming the data we used for each season's prediction we are assuming closely matches what the WHO had available at the time of their vaccine selections.

Once the data ranges for the individual seasons were established, DASH was launched using the run_DASH.pl script. Analysis of the results focused on the HA1 region of HA, as this was the most commonly used subsequence for influenza surveillance used by the WHO. DASH Surveillance candidates from the uncovered clusters were selected if at least three candidates existed in the cluster and the strains were different (often resequenced strains or egg/cell passage variants showed up in the same uncovered clusters). The centroid virus from these clusters was always selected. DASH vaccine candidates were selected from the summarized antigenic coverage prediction results at an antigenic distance cutoff of two (Demonstrated via an AD plot, Figure 16). Antigenic coverage data were sorted by the "in" median results from bootstrapping, followed by the observed "in" from the initial phenotype mapping, then the number of available AD values, and finally by the lower bound of the "in" bootstrap confidence interval. All values were sorted from largest to smallest. Once the data were sorted, the top-scoring candidate was selected as the DASH candidate for each IAV subtype and IBV lineage examined during the season of interest.

Once DASH surveillance candidates and vaccine candidates were predicted we compared these results to the manufacturing and vaccine seeds recommended by the WHO for the subject season. To determine concordance between prediction and recommendation, two antigens were considered antigenically like, if the antigenic distance between them is within two AD units. To make this analysis more stringent, a more conservative cutoff using an antigenic distance of one was utilized to determine concordance. If the DASH candidate and the WHO candidate were the same, the prediction

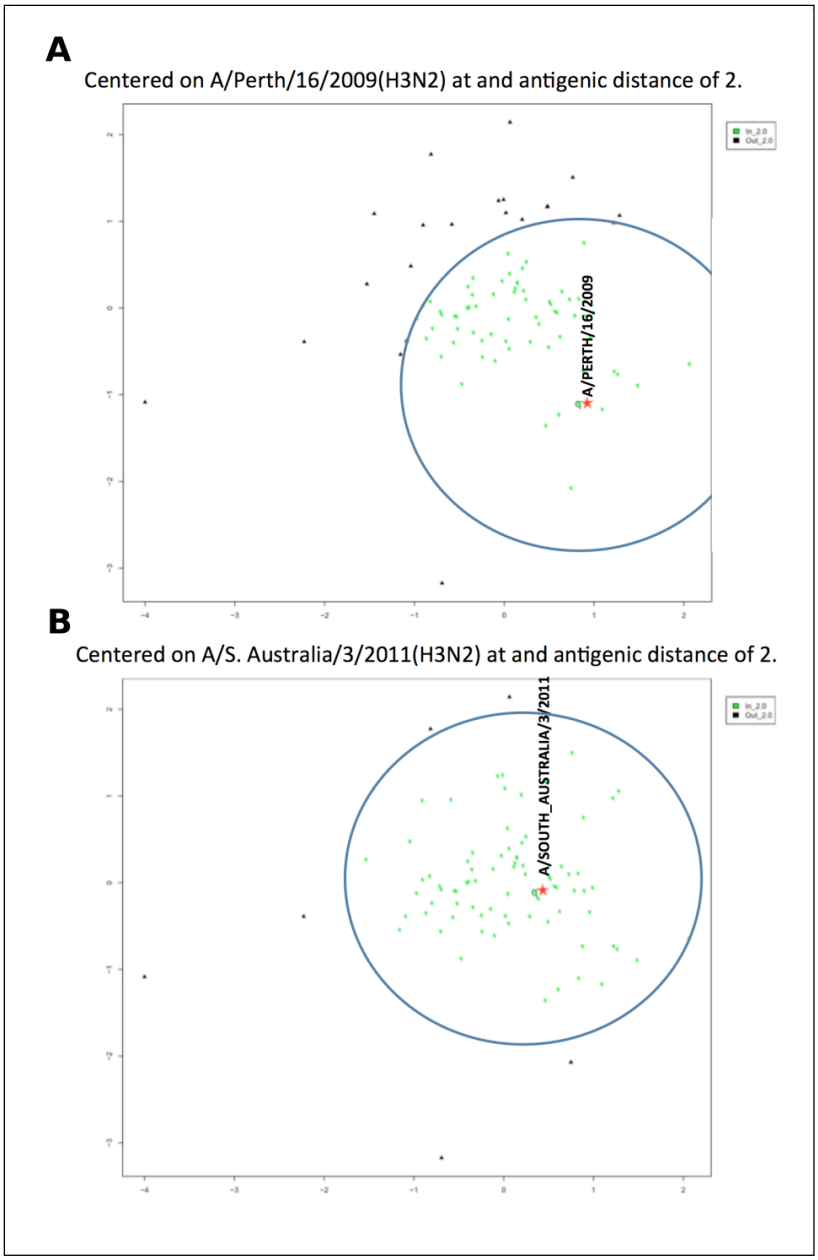


Figure 16 Antigenic Distance Plot HI Data Collected for the 2012-2013 Vaccine Candidate Selection. (**A**) AD plot is centered on A/Perth/16/2009, all green points (within the circle) are viruses considered to be within an antigenic distance of two of this virus. (**B**) Same AD plot except now we have centered the plot on A/S. Australia/3/2011. Comparing these two graphs, A/S. Australia/3/2011 is within an antigenic distance of two of a greater proportion of circulating viruses.

was labeled as exact. If the DASH candidate and the WHO candidate were ≤ 1 AD unit of each other, then they were labeled as concordant. Finally, if the DASH and WHO candidate were > 1 AD unit of each other, then they were labeled as discordant. DASH surveillance candidates were examined to see if any of the candidates matched the manufacturing or vaccine seeds later recommended by the WHO. If a WHO candidate appeared in an uncovered cluster, but not as the centroid, we evaluated how close said candidate was to our surveillance candidate by estimating the quartile rank of the pairwise sequence distance between the two candidates and the remaining distances available for the subject sequence. This resulted in an estimate of how close our predictions were to the WHO candidate viruses. These comparisons were repeated for all the seasons of interest and summary statistics were generated to indicate how successful DASH was at predicting the same or, in our estimation, better vaccine candidates.

Analysis of Effect of JCVI HI Data

For the 2013 Southern Hemisphere predictions we generated our own HI assay data with synthetically constructed viruses rescued using a reverse genetics system. The synthetic viruses were compared using HI assays and AD data generated using our antigenic distancing pipeline. DASH analyses (with and without additional in-house HI assay data) were performed for IAV H3N2 and H1N1pdm subtypes and for IBV Victoria and Yamagata lineages. All sequences were kept identical, only the addition of JCVI HI data was changed between the DASH runs. DASH dendrograms were cut to generate 10 uncovered protein clusters – that is protein clusters containing viral HA sequences but lacking any HI assay data. To establish measurable

differences between the DASH runs with and without JCVI HI assay data, three measures were used. The height in k of the dendrogram at which 10 uncovered clusters were identified was recorded. The median and range of the uncovered cluster sizes were also recorded and compared.

Results

Preliminary Data Analyses

Several preliminary experiments were conducted to establish the feasibility of later work. These experiments included a reanalysis of the 2004 *Smith et al.* data set (42). The results uncovered patterns in the data that refinements in the analysis methodology could potentially exploit, for example, the discovery of a close relationship between Ward's protein clusters and the *Smith et al.* antigenic groups.

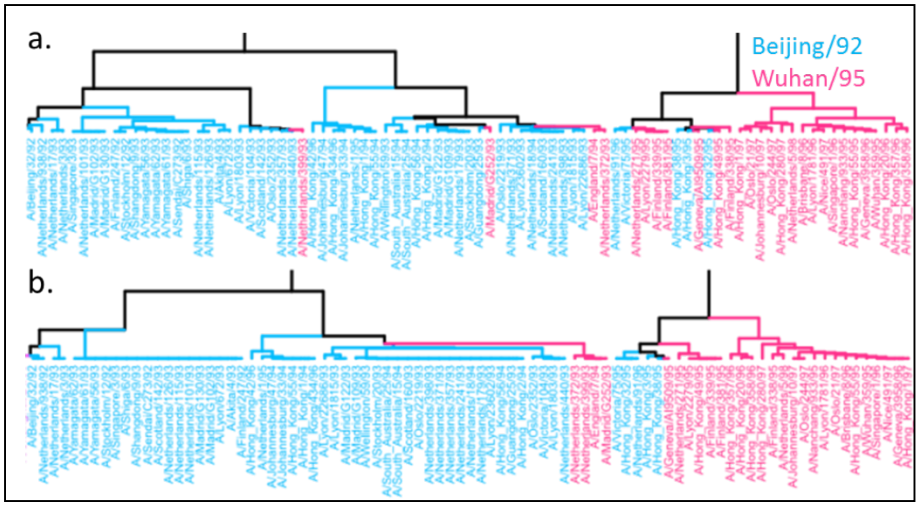


Figure 17 Protein Clusters of HA Sequences From Smith *et. al.* 2004. (**A**) With no protein masking some sequences interdigitate into neighboring antigenic groups. (**B**) Clustering on known epitopes A and B only, the mis-grouped viruses group together.

***Analysis of the Smith et al. 2004 Data set* –** Early trials with mapping antigenic traits onto dendrograms generated through protein clustering were conducted using a data set first described by *Smith et al.* in 2004 (42) when demonstrating antigenic

cartography was possible. This data set includes antigenic groups from H3N2 viruses from 1968 until 2003. The antigenic assignments from this paper were mapped onto dendrograms generated using the Ward's algorithm (74), among others, although the coloring of the leaves of the dendrograms was done according to the antigenic group assignments laid out in the data set (Figure 17). These analyses showed that it should be possible to cluster proteins and have fairly consistent clustering according to antigenic group. The protein masking described above was also applied to these data to see if different treatments would increase the efficiency of clustering. In most cases, clustering was improved using known epitope mask data; however, not all individual known epitopes cleaned up the cluster antigenic group assignments with the equal accuracy. Two problems arose from these analyses. First, some cluster assignments were confused by single amino acid changes in specific viruses across clusters. A sequence could maintain the overall profile of its cluster, but could be colored with a neighboring antigenic group. In nine cases where this was observed, the mis-grouping was the result of one amino acid change (Figure 17). This problem was noted in the *Smith et al.* 2004 paper (42). Our methods may be susceptible to clustering problems if single amino acid changes can alter the antigenicity of a protein and we are using a mask that does not heavily weight those positions. The second issue we discovered was that it may be difficult to find a height cutoff that properly cuts all clusters accurately with respect to antigenicity. This may be due to uneven rates of evolution in various protein clusters that lead to higher amounts of sequence variation with little antigenic change or vice versa. Others have

documented the tendency for antigenic drift to follow a punctuated pattern and sequence evolution to have a more uniform rate (42, 65).

Development of Distancing of Antigenicity by Sequence-based Hierarchical Clustering (DASH) – As a natural extension of the work performed above, we developed a suite of software called Distancing of Antigenicity by Sequence-based Hierarchical Clustering (DASH). Our early work with DASH was focused on contemporary

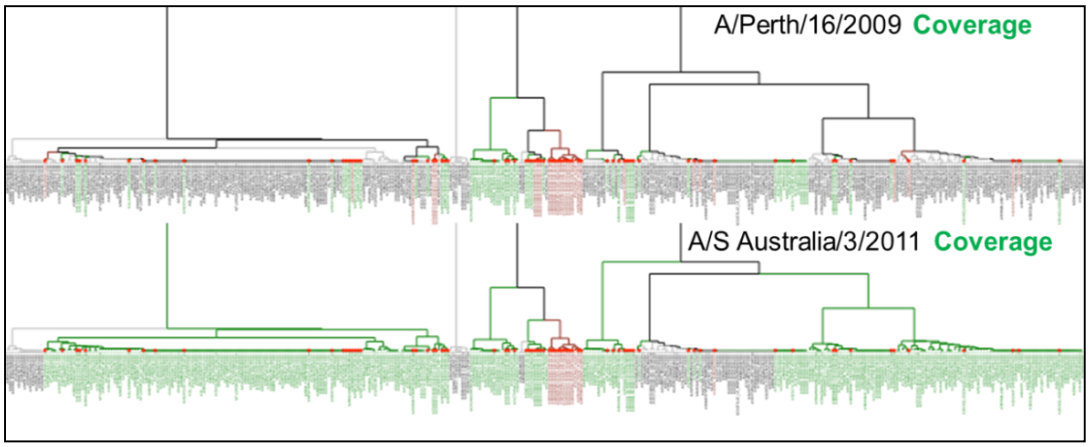


Figure 18 Preliminary Analysis with DASH's Antigenic Coverage Prediction. A comparison of the top and bottom plots indicate that the H3N2 A/S. Australia/3/2011 virus is predicted to cover (green leaves in tree) a much higher percentage of viruses compared to the old vaccine strain A/Perth/16/2009.

sequence data from 2012, with one notable exception. For the vaccine candidate virus selection for the 2012-2013 Northern Hemisphere (NH) influenza season, we have accurately predicted the need for, changes in the IAV H3N2 and the IBV virus vaccine seeds. The H3N2 vaccine seed that we predicted is very close in sequence space to the actual candidate, and we were able to identify the protein cluster from which the H3N2 vaccine was selected. Our centroid virus from that cluster was not the candidate selected, but it did match the vaccine seed we predicted as the best for covering the antigen using the phenotype mapping analysis of DASH. It should be noted that the actual vaccine seed selected for the 2012-2013 vaccine was not

analyzed in any HI experiment publically available; however, the antigen our algorithm predicted was the closest to the actual recommended virus of any for which HI data was available (Figure 18).

In addition to the analysis for 2012-2013 season, the first phase analysis (focusing on DASH surveillance candidates) was performed on H3N2 viruses circulating one year prior to the vaccine candidate selection of 2003 for the NH season 2003-2004. It should be noted that the 2003-2004 season is a known year where the selection and surveillance system did not accurately select the correct vaccine seed (70). In this analysis we were able to identify MAR candidates from seven clusters, including one cluster which contained the antigen selection A/Fujian/411/2002. This virus was not easily manufactured. A centroid virus from another cluster, A/Wyoming/3/2003, was eventually used as the vaccine seed as a result of a subsequent selection meeting (Figure 19). It is unclear at this time if the success in predicting this is due to having a more complete picture of the sequences circulating at the time, or if the success is based on the robustness of

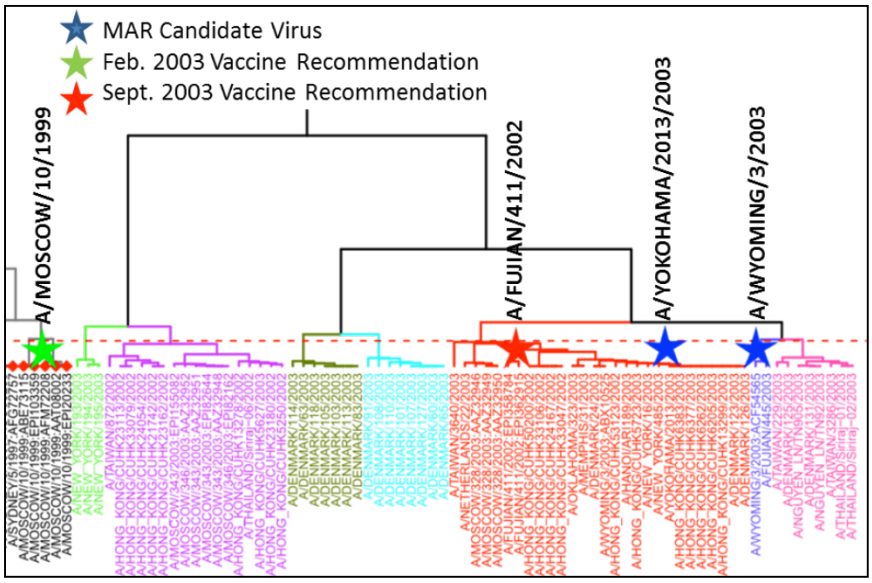


Figure 19 Analysis of the 2003-2004 Vaccine Strain Selection. This uncovered clusters plot suggests that DASH would have identified several MAR candidates similar to A/Fujian/411/2002, including A/Wyoming/3/2003, which was selected for the H3N2 vaccine at a subsequent meeting.

the analyses presented here. It is clear, however, that all viruses isolated from humans should be rapidly sequenced and those data released to the public (75), as the methods described here may be able to identify important variations in influenza far faster than current methods.

Retrospective Analyses Compared DASH Predictions with WHO Influenza Vaccine Strain Recommendations.

To evaluate the ability of DASH to predict vaccine strains, a retrospective analysis was performed for each human seasonal IAV subtype and IBV lineage beginning in 2002, except for pandemic H1N1 (H1N1pdm), which was analyzed since its introduction into humans in 2009. For each NH and Southern Hemisphere (SH) influenza season, DASH was used to predict the vaccine strains that provided the greatest predicted percent antigenic coverage against circulating viruses. If the DASH-predicted strain matched either the WHO-recommended vaccine strain or an antigenically similar strain (defined as ≤ 1 antigenic distance from the WHO-recommended vaccine strain), then the concordance was considered exact. For seasons where the vaccine strains were generally well-matched against circulating viruses based on post-season vaccine effectiveness data, we expected a high concordance between the DASH predictions and the WHO recommendations. In seasons with a poorly matched vaccine strain, we evaluated whether our DASH-predicted strain may have provided a better-matched vaccine than the WHO-recommended strain. Overall, the results of this retrospective comparison indicated a high level of concordance between the vaccine strain identified by DASH analysis and the WHO recommendation. For IAVs H3N2, H1N1, H1N1pdm, IBV Yamagata,

and IBV Victoria, the overall concordance rates were 77.3% (17/22), 100% (15/15), 100% (7/7), 100% (5/5) and 85.7% (12/14), respectively (Table 2). However, using an antigenic distance of < 2 (the accepted upper bound for antigenic similarity) yielded a concordance of 95% (21/22) for H3N2.

For the NH season of 2003-2004, the H3N2 DASH analysis selected A/Fujian/411/2002 as the vaccine candidate, which was in disagreement with the WHO recommendation of A/Moscow/10/1999 (Table 2). During the 2003-2004 season, H3N2 viruses predominated and caused a relatively severe influenza epidemic (76). The WHO-recommended vaccine A/Moscow/10/1999 had only modest vaccine effectiveness, estimated to be 47% (77), and it was subsequently replaced with A/Fujian/411/2002 in the next season (Table 2). This demonstrates that our DASH analysis was able to identify the need to change to A/Fujian/411/2002 one season in advance of the updated WHO recommendation. The DASH analysis predicted A/Fujian/411/2002 because of its higher predicted coverage against circulating strains ("in" = 26.1%) compared with A/Moscow/1999 (estimated coverage of only 3.1%). Because there were limited HI assay data available for that year's circulating viruses, the majority of virus coverage (96.9%) was found to be "unknown" by DASH.

Table 2 Results of a Retrospective Analysis of DASH Candidate Predictions Compared to WHO Vaccine Selections for 22 Influenza Seasons Between 2002 and 2013.

		WHO Recommendation			DASH Recommendation				
Influenza			Predicted Coverage			Predicted Coverage		Antigenic Distance from	
Season	Hemisphere	Strain Name	In	Unknown	Strain Name	In	Unknown	WHO Recommendation	Concordance
2013	Southern	A/Victoria/361/2011	47.4%	49.1%	A/Brisbane/299/2011	84.4%	14.8%	0.40	Like
2012-2013	Northern	**A/Victoria/361/2011**	59.5%	35.9%	A/South_Australia/3/2011†	--	--	--	Exact
2012	Southern	A/Perth/16/2009	24.7%	53.4%	A/Rhode_Island/1/2010	67.8%	30.7%	1.09	Discordant
2011-2012	Northern	A/Perth/16/2009	30.9%	50.9%	A/Perth/10/2010	73.4%	23.4%	1.18	Discordant
2011	Southern	A/Perth/16/2009	45.3%	40.6%	A/Hong_Kong/34430/2009	54.4%	36.2%	0.97	Like
2010-2011	Northern	A/Perth/16/2009	55.4%	17.2%	A/Hong_Kong/1985/2009	62.4%	11.9%	0.37	Like
2010	Southern	**A/Perth/16/2009**	44.8%	11.3%	A/Perth/16/2009	--	--	--	Exact
2009-2010	Northern	A/Brisbane/10/2007	27.5%	62.8%	A/Sweden/3/2008	82.9%	16.7%	0.39	Like
2009	Southern	A/Brisbane/10/2007	32.2%	66.9%	A/Texas/37/2007	53.5%	45.4%	0.21	Like
2008-2009	Northern	A/Brisbane/10/2007	43.1%	55.0%	A/Brisbane/10/2007	--	--	--	Exact
2008	Southern	**A/Brisbane/10/2007**	37.0%	62.8%	A/Wisconsin/3/2007‡	34.6%	62.8%	1.38	Discordant
2007-2008	Northern	A/Wisconsin/67/2005	3.9%	68.4%	A/Hong_Kong/4443/2005	9.2%	84.6%	1.74	Discordant
2007	Southern	A/Wisconsin/67/2005	4.8%	88.4%	A/Wisconsin/67/2005	--	--	--	Exact
2006-2007	Northern	**A/Wisconsin/67/2005**	98.5%	1.5%	A/Wisconsin/67/2005	--	--	--	Exact
2006	Southern	A/California/7/2004	51.6%	46.9%	A/California/7/2004	--	--	--	Exact
2005-2006	Northern	**A/California/7/2004**	81.2%	16.8%	A/California/7/2004	--	--	--	Exact
2005	Southern	**A/Wellington/1/2004**	14.8%	84.2%	A/Wellington/1/2004	--	--	--	Exact
2004-2005	Northern	A/Fujian/411/2002	96.5%	3.5%	A/Fujian/411/2002	--	--	--	Exact
2004	Southern	**A/Fujian/411/2002**	25.0%	73.1%	A/Fujian/411/2002	--	--	--	Exact
2003-2004	Northern	A/Moscow/10/1999	3.1%	96.9%	A/Fujian/411/2002	26.1%	72.0%	3.34	Discordant
2003	Southern	A/Moscow/10/1999	92.3%	7.7%	A/Moscow/10/1999	--	--	--	Exact
2002-2003	Northern	A/Moscow/10/1999	100.0%	0.0%	A/Moscow/10/1999	--	--	--	Exact
								Overall Concordance:	**77.3%**

† The sequence and HI assay data for A/Victoria/361/2011 was not publically released until after the vaccine recommendation was made. The HA segment for A/South_Australia/3/2011 only differed from A/Victoria/361/2011 in the signal peptide.

‡ A/Brisbane/10/2007 was the DASH recommendation when the requirement for ≥ 40 antigenic distance measurements was relaxed.

Due to the limited activity in past few years of the seasonal H1N1 to a large naïve population of hosts to H1N1pdm, these viruses have been found to have lower diversity compared with H3N2 and IBV. This was reflected in our analyses. DASH results had maximal concordances with WHO-selected vaccine strains for seasonal H1N1 (100%, 15/15) and H1N1pdm (100%, 7/7). For H1N1pdm, the DASH-selected strains were essentially antigenically identical to the WHO recommendations, with computed antigenic distances of < 0.1. Differences in the predicted coverage of circulating strains between the WHO recommendations and the DASH predictions are likely due to the conservative inference rules of DASH. The predicted coverage of DASH-recommended H1N1pdm strains against circulating viral strains was supported by more HI assay evidence than for the WHO recommendations; therefore, a larger proportion of circulating viral sequences were assigned to "unknown" coverage for the WHO vaccine strain recommendations.

DASH predictions had 100% (5/5) and 85.7% (12/14) concordances with WHO vaccine strains for the IBV Yamagata and Victoria lineages, respectively. Since the 2002-2003 Northern Hemisphere season, the WHO recommendation has essentially alternated lineages from Victoria to Yamagata and back twice, and recommendations have only changed when the lineage has changed. That means that, within each lineage's succession, the WHO's recommendation has remained constant. This suggests the anticipation of lineage swapping and the aim to not burden the manufacturers with a new strain that would only be utilized for one season. This may explain the discordant prediction in the 2007-2008 Northern Hemisphere season, where the predicted coverage by the WHO-recommended Victoria strain (B/Malaysia/2506/2004) was only 27.5%, whereas the predicted coverage by the DASH-recommended Victoria strain (B/Victoria/304/2006) was 100%. The relatively low prevalence of waning IBV lineages (29% on average for all IBV (78)) in human population is likely a mitigating factor for not matching the most dominant genotypes.

For the cases where the predicted coverage for DASH-recommended strains exceeded that for WHO-recommended strains, it is difficult to evaluate whether the DASH candidate would have had a better vaccine efficacy without controlled experiments. Even correlating the predicted coverage of the WHO recommendations with vaccine effectiveness data is difficult despite a number of reports on vaccine effectiveness (77, 79-81). These reports have largely been inconsistent in their goals and methodologies, leading to variations in their criteria for determining vaccine effectiveness and efficacy. Furthermore, point estimates for effectiveness are also

buttressed with wide and overlapping confidence intervals and may be estimated based on models with inconsistent covariates controlled between analyses.

DASH Identified Drift Variants for Additional Testing and Ranked Potential Vaccine Candidates

In addition to predicting coverage and selecting a vaccine strain, DASH identifies circulating drift variants for which there is no, minimal, or conflicting antigenic data. DASH chooses candidates from among these variants with the aim of increasing coverage knowledge. These DASH surveillance candidates can then be rescued using synthetic genomics and reverse genetics, used to generate ferret anti-sera, and analyzed further using the HI assay and the newly generated anti-sera. This data can be fed back into DASH to refine its vaccine prediction. In many cases, DASH-selected candidates, or very closely related strains, have become the WHO-recommended vaccine strains or vaccine seed stocks in upcoming seasons (as explained below and shown in Table 3).

Table 3 WHO Influenza Vaccine Recommendations and Accepted Seeds Viruses Compared to Time Similar or Identical Viruses Were Identified as DASH Surveillance Candidates.

Influenza Season	Hemisphere	WHO Recommendations		DASH
		Vaccine Strains	Accepted Seed Stocks	STICK Candidates
2013-2014	Northern	A/Victoria/361/2011 (cell propagated)	A/Texas/50/2012	
2013	Southern	A/Victoria/361/2011	A/Ohio/2/2012 A/Maryland/2/2012 A/South_Australia/30/2012 A/Brisbane/1/2012 A/Brisbane/6/2012	
2012-2013	Northern	**A/Victoria/361/2011**		A/Almaty/3277/2012 [α]
2012	Southern	A/Perth/16/2009		
2011-2012	Northern	A/Perth/16/2009		
2011	Southern	A/Perth/16/2009	A/Wisconsin/15/2009 A/Victoria/210/2009	
2010-2011	Northern	A/Perth/16/2009	A/Wisconsin/15/2009	
2010	Southern	**A/Perth/16/2009**		
2009-2010	Northern	A/Brisbane/10/2007	A/Uruguay/716/2007	A/Victoria/210/2009 A/New_York/3148/2009 [β]
2009	Southern	A/Brisbane/10/2007	A/Uruguay/716/2007	
2008-2009	Northern	A/Brisbane/10/2007		
2008	Southern	**A/Brisbane/10/2007**		
2007-2008	Northern	A/Wisconsin/67/2005	A/Hiroshima/52/2005	A/Brisbane/10/2007 A/Uruguay/716/2007
2007	Southern	A/Wisconsin/67/2005	A/Hiroshima/52/2005	
2006-2007	Northern	**A/Wisconsin/67/2005**	A/Hiroshima/52/2005	
2006	Southern	A/California/7/2004	A/New_York/55/2004	A/Wisconsin/67/2005
2005-2006	Northern	**A/California/7/2004**	A/New_York/55/2004	A/Wisconsin/67/2005
2005	Southern	**A/Wellington/1/2004**		A/New_York/55/2004 A/Waikato/1/2004 [γ]
2004-2005	Northern	A/Fujian/411/2002	A/Wyoming/3/2003 A/Kumamoto/102/2002	
2004	Southern	**A/Fujian/411/2002**	A/Wyoming/3/2003 A/Kumamoto/102/2002	
2003-2004	Northern	A/Moscow/10/1999	A/Panama/2007/1999	
2003	Southern	A/Moscow/10/1999	A/Panama/2007/1999	
2002-2003	Northern	A/Moscow/10/1999	A/Panama/2007/1999	A/Fujian/411/2002

[α] Shared cluster with A/Texas/50/2012 (0.10 percentile) and A/Brisbane/1/2012 (13.30 percentile)
[β] Shared cluster with A/Wisconsin/15/2009 (0.62 percentile)
[γ] Shared cluster with A/California/7/2004 (0.51 percentile)

DASH-selected candidates for additional testing and analysis are identified for each IAV subtype and IBV lineage by first deciding the maximum number of viral strains to select (n_{VS}) and then iteratively cutting the tree towards the leaves until the number of clusters without any antigenic information equals n_{VS}. Clusters with only 1 member are not selected because they are considered spurious. A strain closest to the centroid of each identified cluster is selected based on an analysis of molecular variance (AMOVA) (82). In the H3N2 analysis presented here, n_{VS} was set *a priori* to 10.

A DASH analysis of H3N2 viruses identified A/Fujian/411/2002 as a DASH-selected candidate during the 2002-2003 NH season, three seasons before it became the WHO-recommended vaccine strain in the 2004 SH season (Table 3). During the 2005-2006 NH season, A/Wisconsin/67/2005 was identified two seasons before the WHO recommended it as a vaccine strain. Importantly, by the second time A/Wisconsin/67/2005 was chosen as a DASH-selected candidate for the 2006 SH season, HI assay data for this strain and its anti-sera were still not available in the public databases, despite its significant difference from the majority of the sampled strains. Three DASH-selected candidates (A/Almaty/3277/2012, A/New York/3148/2009, and A/Waikato/1/2004) shared the same cluster with a WHO-recommended vaccine strain or accepted vaccine seed stock. Although these isolates were different than the WHO-recommended strains, their sequence similarity was high enough to consider them as antigenically similar strains.

However, in hierarchical clustering, the number of members sharing the same cluster depends on the height at which the tree is cut. Thus, it is misleading to simply claim that a DASH vaccine candidate is in the same cluster as a WHO-recommended strain. In order to evaluate the probability of randomly choosing a strain by chance that was more similar to the targeted strain, we sorted all the strains available in the season of interest by decreasing sequence similarity relative to the target strain and then computed the percentile of the strains that had been chosen based on that ordering. For example, A/Almaty/3277/2012 shared a cluster with A/Texas/50/2012 and was in the top 0.1 percentile of all strains in the 2012-2013 season. In other words, of all the strains analyzed in the given season, 100 - 0.1% = 99.9% of strains

were less similar and therefore were less preferred proxies for A/Texas/50/2012 than was A/Almaty/3277/2012. Equally, highly similar DASH-selected candidates (in the top 1 percentile) were also identified for A/Wisconsin/15/2009 (in the top 0.62 percentile) and A/California/7/2004 (in the top 0.51 percentile).

Additional HI Assay Data from DASH-Selected Synthetically Generated Viruses Improved DASH Vaccine Strain Prediction

The evaluation of the impact of DASH-directed surveillance showed that ,in all cases, JCVI HI assay data increased the proportion of circulating viruses with antigenic coverage (Table 4).

Table 4 Evaluation of The Impact of DASH Directed Antigenic Surveillance During the 2013 Southern Hemisphere Vaccine Selection.

Type	Data set	Tree Height	Median (Range)	% Uncovered
H3N2	w/JCVI	36	7.5 (2,17)	9.44% (81/858)
	w/out	35	9.5 (2,17)	10.37% (89/858)
	Difference	-1	2	0.93%
H1N1pdm	w/JCVI	25	7.5 (4,29)	21.60% (108/500)
	w/out	19	12.5 (1,50)	31.60% (158/500)
	Difference	-6	5	10.00%
IBV Vic	w/JCVI	23	4.5 (1,18)	23.83% (66/277)
	w/out	17	5 (1,18)	28.52% (79/277)
	Difference	-6	0.5	4.69%
IBV Yam	w/JCVI	24	2.5 (1,6)	17.71% (31/175)
	w/out	18	2.5 (1,11)	23.43% (41/175)
	Difference	-6	0	5.71%

Predicted antigenic coverage of surveillance sequences for H3N2 was improved by 0.93% with the inclusion of JCVI HI assay data. The tree height required to identify 10 uncovered clusters increased by one (meaning a deeper cut into the tree), and the median uncovered cluster size decreased by two. The predicted antigenic coverage for H1N1pdm improved by 10.00% with the inclusion of JCVI HI assay data. The tree height required to identify 10 uncovered clusters increased by six, and the median

uncovered cluster size decreased by five. The predicted antigenic coverage by the Victoria and Yamagata lineages of IBV improved 4.69% and 5.71%, respectively, with the inclusion of JCVI HI assay data. The tree height required to identify 10 uncovered clusters increased by six in both cases, and the median uncovered cluster size decreased by 0.5 for the Victoria lineage. The median uncovered cluster size for the Yamagata lineage did not change, although the range noticeably shifted toward smaller cluster sizes.

The overall impact of these analyses on potential vaccine candidate viruses depends in part on how many antigens overlap between HI assays and how many new DASH-selected candidate viruses are tested, as well as on the overall distribution of tested viruses across the protein cluster dendrogram. This analysis is best viewed as a method for increasing the overall knowledge about a particular antigen and about circulating viruses in general. A candidate may look more or less desirable as a vaccine strain, but the net gain is an improved resolution of the antigenic phenotype across all circulating viruses. The results described below demonstrate that targeted antigenic analysis through HI assay testing of DASH-selected candidates can have a significant impact on the resolution of the antigenic phenotype. This information is highly relevant for vaccine production.

A total of 81 viruses were tested in the two CDC H3N2 HI assays. JCVI HI assay testing of DASH-selected candidates provided data for an additional 10 viruses, as well as providing additional HI assay data for nine CDC-tested viruses. Changes in the percentage of "in" predictions after including the additional JCVI HI assay data ranged from -16.89% to 25.32%, whereas changes in the percentage of "out"

predictions ranged from -46.12% to 2.64%, and the percentage of "unknown" predictions ranged from -10.90% to 30.86% (Table 5 and Figures 20 and 21).

Table 5 Relative Shift in Observed Predicted Antigenic Coverage for Nine H3N2 Viruses With and Without JCVI HI Data.

Virus	Shift In	Shift Out	Shift Conflicted	Additional In	Additional Out	Original Rank	New Rank	Rank Change
A/Hong Kong/4913/2011	25.32%	-14.42%	-10.90%	17	3	47	2	45
A/Victoria/208/2009	15.26%	-46.12%	30.86%	18	2	80	75	5
A/Bangladesh/5071/2011	8.47%	-0.53%	-7.93%	17	3	72	72	0
A/Perth/16/2009	5.65%	-2.21%	-3.43%	16	4	77	76	1
A/Berlin/3/2012	2.64%	-2.65%	0.01%	17	3	60	56	4
A/Alabama/5/2010	2.64%	-2.65%	0.01%	17	3	63	57	6
A/South Australia/3/2011	2.38%	0.04%	-2.41%	16	4	6	1	5
A/Hong Kong/3969/2011	-1.99%	-2.61%	4.60%	17	3	22	23	-1
A/Brisbane/299/2011 Egg	-16.89%	2.64%	14.25%	6	14	5	22	-17

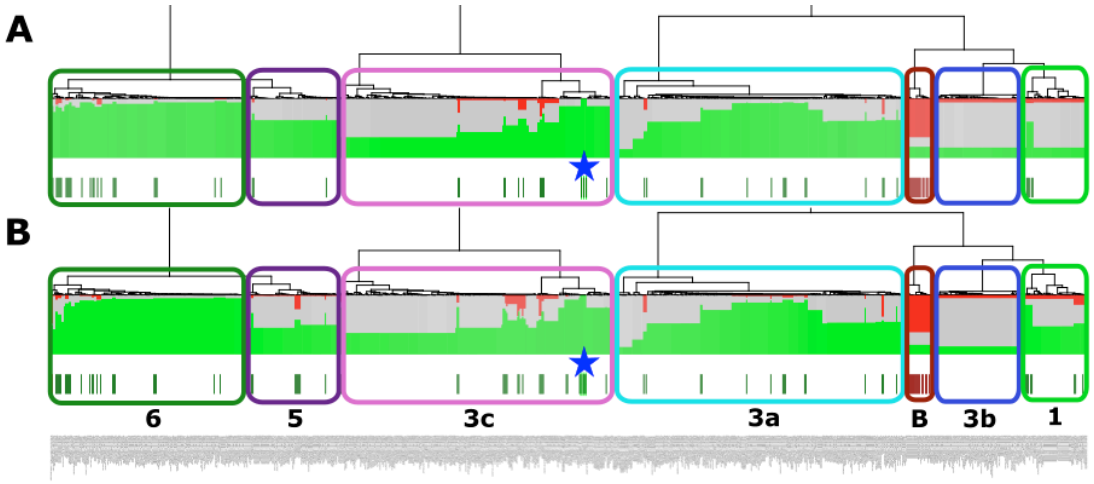


Figure 20 DASH Diagrams Comparing Antigenic Coverage Prediction for A/S. Australia/3/2011 Without (A) and With (B) JCVI HI Data. The relationship between circulating viruses over the collected between August 2011 and August 2012 is depicted in the dendrogram at the top of each panel. The green bars in the center of each plot indicate the ratio of "in" (green), "out" (red), and "unknown" (grey) predictions for each virus in the dendrogram over 80 bootstrap replicates. The green and red bars in the bottom row of each plot indicate which viruses in the dedrogram also had HI assay data. Overall these diagrams are very similar. A/S. Australia/3/2011 is indicated with the blue star. Phylogenetic clades as defined by phylogenetic analysis and comparisons to WHO nomenclature are indicated with colored boxes and labels at the bottom of the second panel. These diagrams can be considered to be very similar, however **(B)** shows slightly more predicted coverage around viruses with new HI data in clades 3c and 1. These differences move A/S. Australia/3/2011 up in rank from six to one of all viruses tested.

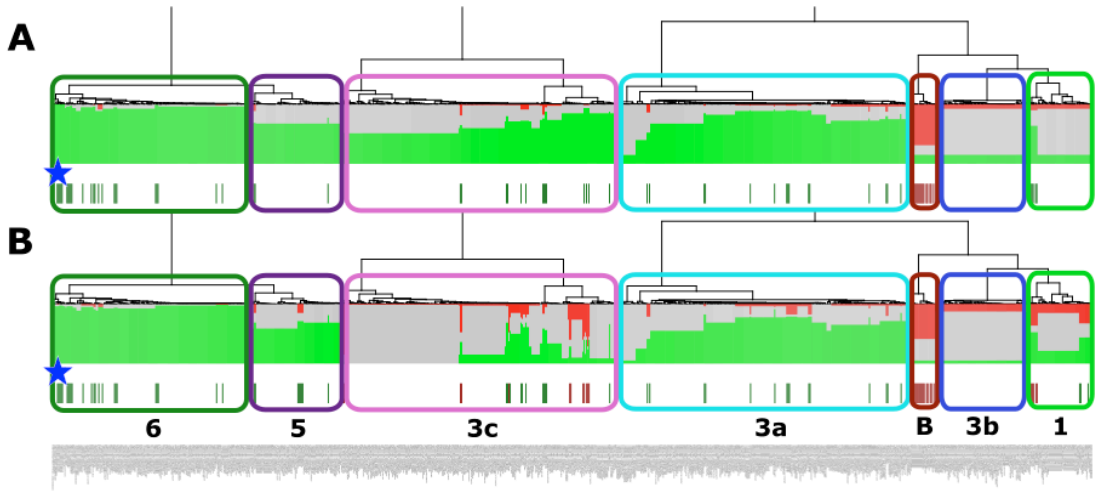


Figure 21 DASH Diagrams Comparing Antigenic Coverage Prediction for A/Brisbane/299/2011 Without (A) and With (B) JCVI HI Data. The differences between these two diagrams show a large drop in predicted antigenic coverage of this virus against the 3c clade. Phylogenetic and historical analyses reveal clade 6 to be older than clade 3c. This may indicate the waning protective influence of clade 6 viruses versus clade 3c viruses. PTP analysis reveals clade 3c to be growing. Taken together this indicates A/s. Australia/3/2011 to be a better potential vaccine candidate than /Brisbane/299/2011.

During the 2012-2013 season, it was widely reported that the chosen vaccine strain (A/Victoria/361/2011) underwent significant antigenic changes during its adaptation to growth in eggs (Figure 22). Although no A/Victoria/361/2011 virus in any propagation media was included in the publically available data at the time of its selection as a vaccine recommendation, our data support an antigenic difference between egg- and cell-propagated H3N2 vaccine strains as early as August 2012, which likely contributed to the reduced vaccine effectiveness.

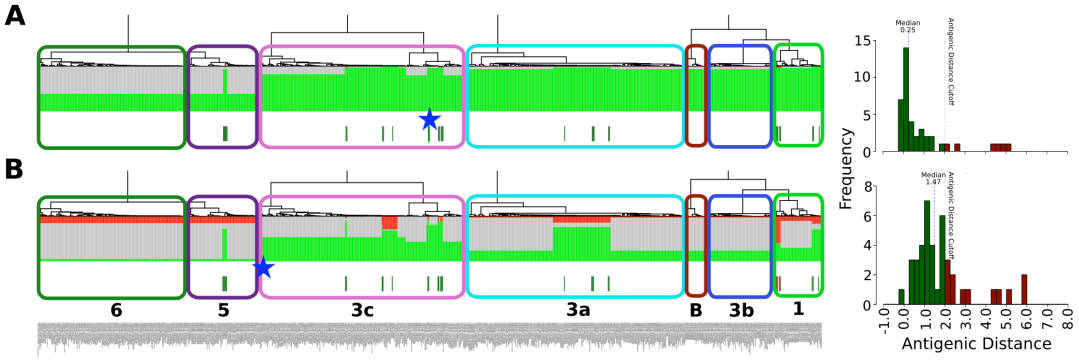


Figure 22 DASH Diagrams Comparing Antigenic Coverage Prediction for A/Victoria/361/2011 Cell (A) and Egg (B) Passaged Virus HI Data With AD Histogram Plots. A comparison of the antigenic coverage

prediction of the cell- and egg-passaged A/Victoria/361/2011 virus reveals different coverage prediction patterns on the various circulating clades. The cell-passaged virus matches well against clades 3A, 3B, 3C, and 1, whereas the egg-passaged virus shows decreased coverage against all clades with particularly poor coverage predictions on clades 1, 5 and 6. The histogram plots on the right of each DASH plot indicate the overall distribution of the AD values used for each prediction. The shift of median AD value toward the antigenic cutoff of 2 for the egg-passaged virus explains this discrepancy in coverage predictions.

A total of 32 H1N1pdm viruses were tested in the CDC HI assay. JCVI HI assay testing of DASH-selected candidates provided data for an additional 16 viruses and added additional HI assay data for seven CDC-tested viruses. Shift in "in" predictions ranged from -2.71% to 37.66%. Shift in "out" predictions ranged from -8.68% to 0.56%. Shift in "unknown" predictions ranged from -34.95% to 3.72% (Table 6 and Figure 23).

Table 6 Relative Shift in Observed Predicted Antigenic Coverage for Seven H1N1 Viruses With and Without JCVI HI Data.

Virus	Shift In	Shift Out	Shift Conflicted	Additional In	Additional Out	Original Rank	New Rank	Rank Change
A/Quebec/RV1432/2011	37.66%	-2.71%	-34.95%	54	4	30	9	21
A/Voronezh/1/2011	6.31%	0.34%	-6.65%	49	9	31	31	0
A/St Petersburg/100/2011	6.20%	0.45%	-6.65%	46	12	29	29	0
A/New York/21/2011	4.96%	-8.68%	3.72%	44	14	33	33	0
A/California/52/2011	0.23%	0.23%	-0.45%	52	6	10	10	0
A/Valparaiso/17275/2011	-2.14%	0.23%	1.92%	53	5	4	8	-4
A/California/7/2009	-2.71%	0.56%	2.14%	49	9	11	28	-17

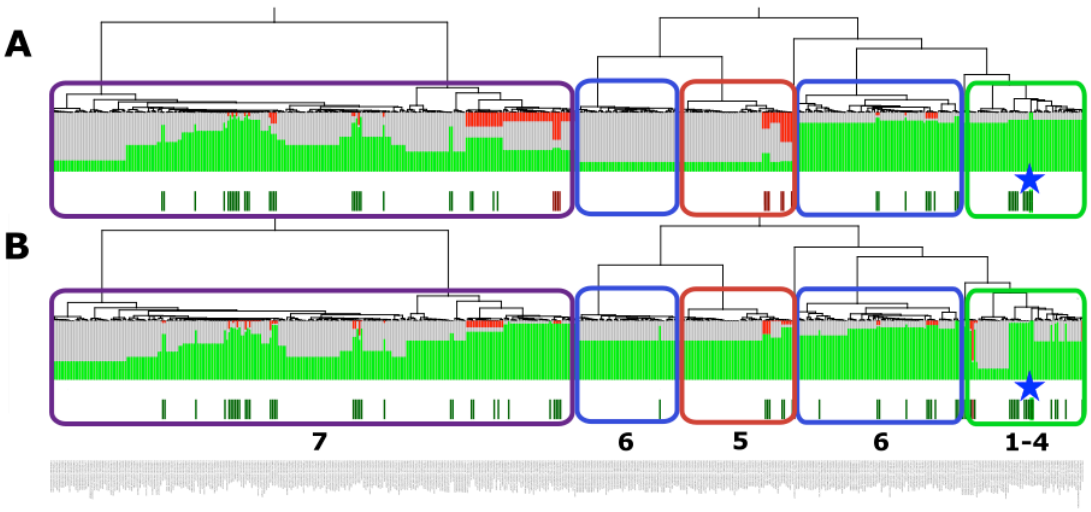


Figure 23 DASH Diagrams Comparing Antigenic Coverage Prediction for A/Quebec/RV1432/2011 Without (A) and With (B) JCVI HI Data. These two DASH diagrams show an increase in predicted antigenic coverage

of circulating H1N1pdm viruses by A/Quebec/RV1432/2011 with the addition of JCVI HI data. Predicted antigenic coverage increased for clades 5, 6 and 7, whereas part of the cluster representing clades 1 through 4 showed decreased predicted antigenic coverage. This virus moved from rank 30 without JCVI HI data to rank nine with JCVI data. This result indicates this virus could be a viable vaccine candidate.

A total of 33 IBV Victoria lineage viruses were tested in the CDC HI assay. JCVI HI assay testing of DASH-selected candidates provided data for an additional eight viruses and added further HI assay data for two CDC-tested viruses. Shift in "in" predictions ranged from 0.45% to 0.59%. Shift in "out" predictions ranged from 0.30% to 0.45%. Shift in "unknown" predictions were measured at -0.89% (Table 7).

Table 7 Relative Shift in Observed Predicted Antigenic Coverage for Two IBV Victoria Lineage Viruses With and Without JCVI HI Data.

Virus	Shift In	Shift Out	Shift Unknown	Additional In	Additional Out	Original Rank	New Rank	Rank Change
B/Nevada/3/2011 Egg	0.59%	0.30%	-0.89%	10	12	9	2	7
B/Kansas/1/2012	0.45%	0.45%	-0.89%	9	13	25	22	3

A total of 27 IBV Yamagata lineage viruses were tested in the CDC HI assay. JCVI HI assay testing of DASH-selected candidates provided data for an additional nine viruses and added further HI assay data for three CDC-tested viruses. Shift in "in" predictions ranged from -14.76% to 19.75%. Shift in "out" predictions ranged from 0.42% to 1.25%. Shift in "unknown" predictions ranged from -20.17% to 13.51% (Table 8 and Figure 24).

Table 8 Relative Shift in Observed Predicted Antigenic Coverage for Three IBV Yamagata Lineage Viruses With and Without JCVI HI Data.

Virus	Shift In	Shift Out	Shift Unknown	Additional In	Additional Out	Original Rank	New Rank	Rank Change
B/New Hampshire/1/2012	19.75%	0.42%	-20.17%	9	13	26	25	1
B/Wisconsin/1/2010	8.52%	0.62%	-9.15%	10	12	24	9	15
B/Finland/39/2010	-14.76%	1.25%	13.51%	8	14	2	10	-8

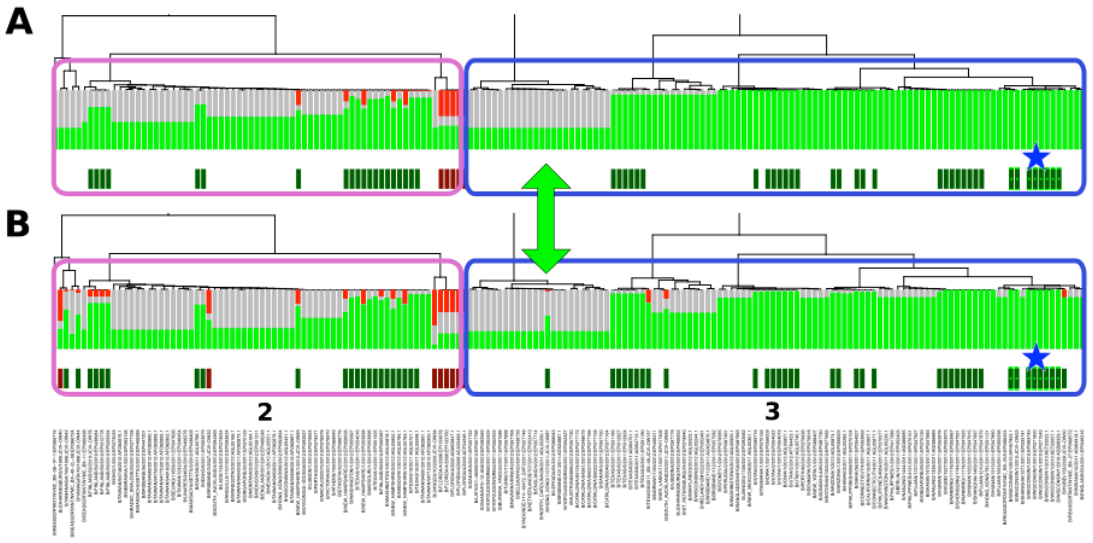


Figure 24 DASH Diagrams Comparing Antigenic Coverage Prediction for B/Wisconsin/1/2010 Without (A) and With (B) JCVI HI Data. These two DASH diagrams show an increase in predicted antigenic coverage of circulating IBV Yamagata lineage viruses by B/Wisconsin/1/2010 with the addition of JCVI HI data, although this change does not seem exceptional in the depiction above. The green double-sided arrow points to one new virus tested by JCVI HI assay. This particular virus contained a HA sequence that was identical to 27 other viruses that were circulating during this time period. DASH applies weights to viruses that are highly represented in the data set. This weighting accounted for an 8.52% increase in predicted antigenic coverage of B/Wisconsin/1/2010 over IBV Yamagata lineage viruses.

Discussion

The goal of this study was to put forth a method to improve the viral surveillance of influenza virus, with the hope of expediting the vaccine selection process while simultaneously making improvements to its accuracy. Although we believe these methods achieve that goal, a shift in focus is likely necessary. Next-generation sequencing is a key component of that shift, where an unbiased sequencing-first approach to surveillance should be used. It is only through an unbiased sequencing philosophy that the epidemiological dynamics of circulating influenza viruses becomes most accurate. This is especially true in light of the high variability of the HI assay data (and related computations), and the accuracy achieved in modern sequencing techniques. The assumption of reliability of these data as an unbiased surveillance resource is a key underpinning of the methods described here.

This study supports the notion that targeted antigenic surveillance after the sequencing of influenza genomes, rather than large-scale HI assay testing prior to sequencing, is a viable alternative. This method would have advantages over the current method in cost and timesaving. A sequencing-first methodology would give researchers the ability to plan HI experiments in targeted areas of the viral protein space, and hence to maximize the increase of phenotypic information gained through HI testing. Furthermore, these methods provide a strong statistical basis for selecting viruses as potential vaccine candidates based on antigenicity alone. We have also provided evidence of the ability of these methods to resolve and visualize differences in antigenicity in egg- versus cell-propagated viruses, as was the case for A/Victoria/361/2011. This capability may allow for smarter selection of vaccine candidates in the future, as poorly matched egg/cell pairs could be screened out as potential vaccine seeds. Similarly, relatively small differences in antigenic distance (0.5-1.5 AD) can be shown to have large impacts on the predicted viability of one candidate virus over another especially in context of newly emerging clades. These differences only become apparent through our bootstrapping techniques and the use of DASH visualizations (See A/Victoria/361/2011 egg/cell example, Figure 23).

DASH and a related methodology, the PTP, have aided in determining the direction that antigenic assays should be taken. These methods thus enable the sequence-first approach discussed above. PTP is capable of tracking which genetic groups are growing or shrinking in prevalence during a given influenza season. As we described earlier, DASH is able to identify gaps in antigenic information. Taken together, these methods can help researchers direct antigenic characterization to the

clusters of viruses where data is limited and to prioritize the study of those influenza viruses that are seen to have high rates of expansion in prevalence. Although smaller proportion groups are still of interest, since emerging antigenic variant influenza viruses can evolve from any sub population – the relative proportion should be taken into account. It can be assumed that genetic groups that show little sign of expansion have likely not yet acquired the mutations required for immune evasion. In designing HI experiments, sera panels should be derived from the breadth of recently circulating genetic groups, although antigens can be focused on areas with limited antigenic data.

HI experiments (and thus the AD calculation) itself are subject to a high degree of variability. This variability is likely derived from the differences in methodology between labs. These include: the use of regents (such as blood cells) from differing donor species (e.g. turkey, guinea pig, etc.), the use of additives (such as oseltamivir) to aid in NA normalization, and the various polyclonal sera used to inhibit hemagglutination. Not only are polyclonal sera derived from different host species (most commonly ferret or sheep), but the use of outbred animals to generate these sera will likely lead to qualitative differences in results. This is because the genetic background of the host can influence the specific antibodies produced when challenged with a novel pathogen. During the AD computation, the serial dilution processes of the HI experiment can lead to quantization errors where larger distances will become increasingly error prone, as small differences in antibody titer will become large differences after many dilutions. For this reason, we believe AD is most accurate at ≤ 2 AD units or 4-fold dilutions in an HI assay. Similarly, low titer values will also affect the reliability of the results as the dynamic range of the assay is

shortened. Although it is our belief that these sorts of errors can be somewhat ameliorated by centralization of the assays and performing technical replicates of each antigen of interest, HI assays will likely not be scalable to the extent that high-throughput sequencing techniques have become.

The methods described herein should be seen as an enhancement to current viral surveillance and vaccine selection methodologies. It is through the rigorous statistical treatment described above that human biases in the vaccine selection process may be eliminated. It is also our belief that the use of synthetic genomics when coupled with computation, will speed the production of selected candidates, enhance the evaluation of new candidates, and allow for the manufacturing at-risk and stockpiling of viruses in real-time during viral epidemics – ready for later production if similar viruses are chosen as the vaccine candidates.

Conclusion

We have described methods for performing viral genotype and phenotype surveillance, as well as for the mapping of antigenic data on protein dendrograms, for the purpose of evaluating novel viruses for use as vaccine components in the annual influenza vaccine. Through an 11-year retrospective analysis we have demonstrated the ability of these methods to accurately select the same or similar viruses as the WHO-recommended components. In some cases, our methods have predicted the necessary changes in the vaccine seed choice by the WHO, and by our computational estimations have selected more suitable vaccine component viruses. DASH has also a demonstrated ability to identify novel antigenic groups through sequence analysis, and could be used as a surveillance technique to identify manufacturing at-risk

vaccine component candidates for the pharmaceutical industry. Finally, we have demonstrated, through the use of JCVI generated HI assay data, that DASH could be used to enhance antigenic surveillance by facilitating directed HI assays of novel viruses, by reducing redundancy in the surveillance process, and by ensuring the maximum information gain and the lowest investment of labor and reagents.

Chapter 3: VirComp: A Novel Method for Viral Comparative Analysis Using Cluster-Based Gene Constellations

Abstract

Emerging and endemic viral pathogens carry an enormous disease burden on human and animal populations. Morbidity, mortality, and economic losses are just a few ways in which disease burden can be calculated. From the recent reemergence of Ebola in West Africa to the annual epidemics of respiratory diseases such as influenza and respiratory syncytial virus (RSV), research into viral diseases is of utmost importance. Next-generation sequencing technologies have altered our ability to perform viral surveillance and directed research into epidemiological and pathogenic phenomenon related to these pathogens. However, handling of these increasingly large data sets is key to making important inferences about viral epidemiological dynamics. Here we describe a novel method, VirComp, for performing genome constellation analysis in a context specific manner. VirComp when paired with OrionPlot (a new visualization application) enables detailed comparative viral genomics. We show the utility of this method for understanding the reassortment and evolutionary processes in influenza virus, for investigating recombination and performing sequencing validation in RSV, and for carrying out variant screening in Ebola virus. This method will aid in the enormous task of describing new viral data sets in light of large-scale genomic sequencing efforts.

Introduction

Emerging and endemic viral diseases are a great threat to human and animal populations around the world. Pathogens such as influenza and RSV cause large disease burdens during their annual epidemic seasons (83, 84). Influenza virus is thought to cause between 250,000 and 500,000 deaths annually with 3 to 5 million cases (84). Similarly, RSV causes 199,000 deaths in children under the age of 5 worldwide and 10,000 deaths in the US in the 65 plus community (85, 86). The vast majority of RSVs disease burden is felt in developing countries (87). The recent reemergence of Ebola virus has caused widespread death and economic calamity throughout West Africa since early 2014 (88). As of this writing the 2014 Ebola virus epidemic has killed 11,261 people in six countries with laboratory confirmed cases in 10 countries on three continents (89). A total of 27,609 cases are suspected (89). With the current pace of genomics and the introduction of new technologies to expedite viral genome sequencing, new methods for effective viral comparative genomics are increasingly needed. Novel methods for analyzing this onslaught of new data will no doubt enhance our understanding of these and other endemic and emerging pathogens.

The most robust method for comparing viral genomes is to perform phylogenetic analysis. However, comparing viral genomes, taking into account variations across multiple constituent genes, becomes intractable with large data sets across several to dozens of gene-specific phylogenies (90). Methods for quickly categorizing and visualizing categorical assignments are useful in solving the difficulty of making tree based comparisons of genes and genomes (91). This type of

analysis is commonly known as genome constellation analysis, or sometimes genotype analysis (49, 91). Constellation analysis has been used extensively in influenza virus and rotavirus studies to uncover and examine the nature of genomic reassortments in segmented viruses (48, 50, 52, 92). In principle these techniques could just as easily be used to study recombination in non-segmented viruses.

Constellation analysis comes in two basic forms. The first method relies on distance metrics based on sequence similarity alone to categorize differences in the genomic content of the viruses. These approaches can be further categorized as one-to-many comparisons (where one subject is compared against many references) (50), and many-to-many comparisons (where an all-versus-all comparison strategy is employed) (51). The second method utilizes phylogenentic reconstructions as the basis for categorizing the viral genes (48, 91). Distance methods tend to be less computationally rigorous and less time consuming than the phylogenetic approaches, especially if maximum likelihood or Bayesian inferences are made. Furthermore many of the methods described in the literature, including the phylogenetic methods, involve less straightforward approaches to making cutoffs between categories (91). These approaches can make it difficult at times to discern, in a quantifiable way, what the differences are between the genotypes or constellations being examined.

Methods

Data Collection and Preparation

All viral sequence data was collected either from GenBank or GISAID's EpiFlu databases. Viral genomes were selected for analysis according to their complete or near completeness. Influenza virus sequence data were separated by

genomic segment into multifasta files, and labeled according to segment name. RSV and Ebola virus genomes were first annotated with VIGOR 3.0 (22). The resultant CDS file was then separated into gene specific multifasta files, and labeled according to gene name. All sequences were identified by their database accession number. Additional accession number to viral strain name mapping files were generated as a way to associate unique sequence identifiers in the gene-specific sequence fasta files to specific viruses during a later step used to aggregate results. Viral strain names were manually reviewed to ensure consistency in species-specific nomenclature, and to enable exact string matches to be made by the constellation processing scripts.

VirComp Constellation Analysis Pipeline

The VirComp (https://github.com/sschobel/vir-comp) constellation analysis pipeline was implemented in the PERL programming language (Figure 25). The pipeline includes a main workflow wrapper script called run_constellation.pl. The input for the pipeline is a two column tab delimited file (tsv) with the segment or gene names in the first column (matching the fasta files prepared previously) and in the order by which that segment or gene-specific clusters should appear from left to right in the resultant visualization. The second column of the tsv file should include a sort order for the resultant pipeline output. This allows a user to specify both the order of the gene columns and the order by which the columns should be sorted. For example, a useful method is to order the columns by gene synteny in the genome, and sort the columns from genes with the most to least diversity. The pipeline auto-detects the fasta files of the segments (or genes), so long as the names in the first column of the input file matches the gene or segment names of the fasta files and the fasta file

names are suffixed with fasta. This script in turn executes six additional data processing applications. The first of these is MAFFT (93). MAFFT is responsible for generating multiple sequence alignments and was selected due to its speed and accuracy. Multiple sequence alignments were output in the ClustalW format (an interleaved multiple sequence alignment file). After MAFFT, run_constellation.pl executes ClustalALN_to_RDistancematrix.pl, which transforms the gene or segment specific multiple sequence alignments into pairwise distance matrices (39). This script is available through the ANDES suite of deep sequencing analysis software (39). The distance between sequence pairs is calculated using the Hamming distance method for nucleotide and protein alignments (94). These distance matrices are then used multiple times to generate gene clusters using farthest neighbor hierarchical clustering and a set of predefined percent identity cutoffs (100%, 99%, 98%, 97.5%, 95% and 90%). The gene clusters are constructed using Partition_Members_byDistanceMatrix.pl, another member of the ANDES suite (39). Once the gene clusters are generated for all segments or genes at a particular cutoff level the pipeline then executes merge_constellations.pl. This script is responsible for utilizing the accession to strain name maps to match gene cluster assignments for each gene or segment to their respective viral strains. The program then outputs a genome constellation of each gene in the viral genome for every viral strain in the map files. The output tsv file is sorted according to gene-specific columns specified in the original pipeline input. For instance if the user specifies a desire to sort on the HA gene first, then the NA gene, then all of the rest of the genes in order of segment number for an influenza virus, then the output will begin with HA and NA cluster 1

viruses. This would be followed by HA cluster 1 and NA cluster 2 viruses if such combinations existed, and this step is repeated until all viral strains have been output. A second constellation tsv file is also outputted by merge_constellations.pl. This output consists of only the set of unique constellations labeled with the count of observations of that constellation in the input data set. This output can be useful to assess the abundance of a particular constellation and makes it easier to summarize especially large constellation analyses. The constellation analysis pipeline produces two additional files as supplementary annotations for the constellation tsv files, produced by merge_constellations.pl. The first is a list of the viral strain names in each of the unique constellations that have been derived from the input set (at the cutoff being examined). The second is the mapping of the viral strain name to the constellation number. Constellations are numbered according to the sort order specified by the user in the gene order input tsv file. These annotation files are generated using annotate_constellations.pl and annotate_constellation_map.pl respectively. Once run_constellation.pl has completed the generation of the gene clusters, the aggregation of the constellations and the annotation of the constellations at a particular cutoff, the pipeline moves on to the next predetermined cutoff until all cutoffs have been examined.

To reduce redundant computational steps the pipeline is designed to run the alignment and the distance matrix generation steps only once per gene, as these will not change when examining the different percent identity cutoffs. However, the cluster generation, the aggregation of the constellation information, and the annotation of constellations must be performed independently for each cutoff. The

pipeline simply passes these files on to the later processing steps as it iterates over the various cutoffs. The run_constellation_analysis.pl script is also capable of detecting the presence of pre-generated alignments and/or distance matrices. The purpose of this feature is twofold. First, this allows the pipeline to skip the computational expense of producing these files if the user already generated them – for example if the user merely wanted to update a viral strain name in a map file. Second, it allows

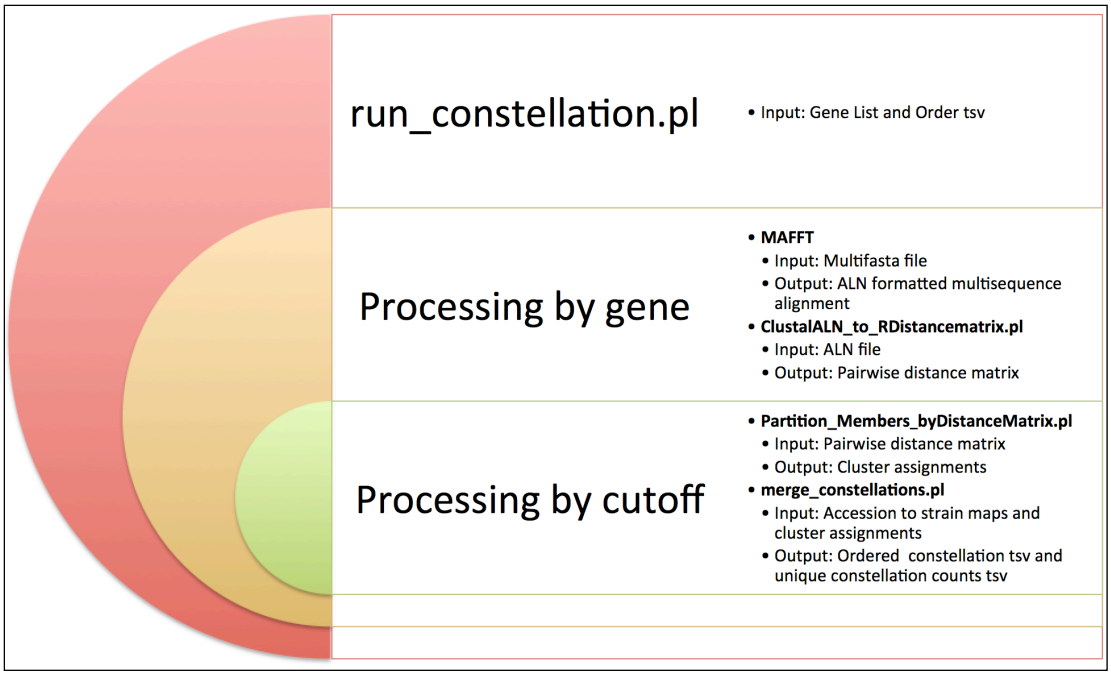


Figure 25 Workflow of The VirComp Constellation Analysis Pipeline. Scripts responsible for various aspects of the computational pipeline are listed next to the processing step. Inputs and outputs of each step of the analysis are also indicated.

the user to provide edited alignments or pairwise distances calculated according to a different algorithm.

Maximum Likelihood Phylogenetics

The H3N2 and RSV data sets were used to generate phylogenies of each viral gene for comparison to the gene-based clustering technique used in this method. The phylogenies were inferred using GARLI Web 2.0 (95) using the general time

reversible substitution model using a discretized gamma distributed site-specific variation model, with four rate categories, and an invariant site parameter.

Sequence Analysis

Sequence analysis for RSV recombination analysis was performed using MAFFT (73) for sequence alignment and CLCBio (96) for alignment visualization. Recombination analysis for confirmation of VirComp results was performed using RDP3 (53).

Constellation Visualization using OrionPlot

OrionPlot (https://github.com/sschobel/orion-plot) is a Java application that processes categorical matrix data sets and generates a concise visual representation from them. The final output is provided in SVG and PNG formats.

The tool provides a multi-platform GUI front-end. This desktop application displays a preview of the output image and allows the user to modify the visualization parameters in real time before generating the final output. A command-line interface is also provided.

Results

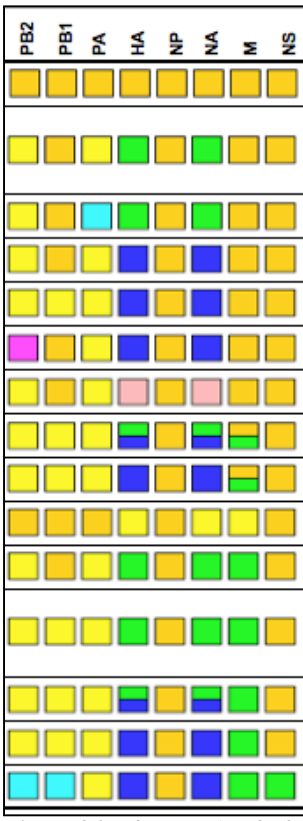
During the development of the VirComp constellation analysis software five separate analyses were performed to assess the utility of this method and to further develop functionality for the software. These analyses consisted of a swine influenza data set, a human seasonal influenza data set, an avian influenza data set, a human RSV data set, and an Ebola virus data set. The former three data sets represent a segmented virus where reassortment of the genomic segments is presumed to be

possible if not commonplace. The RSV data set represents a single-segmented viral genome where reassortment does not occur; however, genomic recombination should be assessed, especially prior to performing any in-depth phylogenetic or other evolutionary based analyses. Finally, the Ebola data set represents a low-variation genome, where the screening of the protein variants can provide support for the design of *in vitro* assays.

Exhibition Swine Influenza Virus Analysis

The swine data set included a set of viral samples from collected from pigs at several county and state fairs in Ohio during the summers of 2009-2011. In addition to the viruses from this study, our analysis included North American swine sequence from the same period, human vaccine virus and H1N1, H1N2, and H3N2 variant viruses (collected as a result of human infections). The cutoff used for this data set was 97.5% (Figure 26). The results of the constellation analysis show a relatively static set of viruses circulating within the Ohio fairs during 2009 and 2010; however, during 2011, a much more dynamic viral landscape emerged. During 2009, two viral constellations were circulating in pigs, one from the H3N2 subtype and one the H1N2 subtype. Similarly, in 2010 just one H3N2 viral constellation was present. In 2011, however fair sampling showed nine distinct viral constellations circulating from both the H3N2 and H1N2 subtypes. Single fairs in 2011 showed the presence of multiple constellations and subtypes. During one fair in particular, sampling showed the presence pigs co-infected with more than one subtype, the hallmark of the reassortment process. During 2011 and 2012, reassorted variant H3N2 virus was found to have infected humans, primarily in the Midwest and associated with the state

fairs. The constellation analysis shows a close relationship between the H3N2 variant sequences found in human children who visited or were in contact with those who had visited county and state fairs. This suggests that swine viruses were being transmitted to humans at these events, and possibly, vice versa. One other interesting feature of the constellation analysis for this study was the switch over from the



Swine				Human		
Count	Subtype	Year	Ohio Fair	Count	Subtype	Year
19	H1N2	2009	A			
24	H3N2	2010	C, D, E	1	H3N2	2010
1	H3N2	2011	Hb			
				1	H3N2	2010
1	H1N2	2011	Hb	1	H1N2	2011
2	H1N2	2011	Hb			
1	H1N2	2011	Hb			
				1	H3N2	2010
7	H3N2/H1N2	2011	Hb			
1	H1N2	2011	C, D, E			
33	H3N2	2009	B, C	1	H3N2	2009
				2	H3N2	2012
35	H3N2	2011	Ha, Hb	8	H3N2	2011
7	H3N2	2011	C, F	77	H3N2	2012
10	H3N2/H1N2	2011	C, D			
48	H1N2	2011	B, G, Hb, C, D			
				1	H1N2	2012

Figure 26 Influenza A H3N2 and H1N2 Genome Constellations Present in Exhibition Swine From Ohio Fairs Between 2009 and 2011. The constellation analysis shows the presence of numerous distinct genome constellations at the 97.5% cutoff in exhibition swine during the time period specified. 2009 and 2010 constellations show no mixed infections or reassortment events. 2011 constellations do show evidence of mixed infections (indicated by squares shaded with two different colors) and reassortment within and between fairs (5).

classical matrix to the pandemic matrix gene (a signature sequence from the 2009 H1N1 pandemic) during the 2011 fair sampling. This suggests a continual mixing of various human and avian origin gene segments within the swine population.

Human H3N2 Influenza from Houston, Texas During the 2012-2013 Season

A.

	Genome Constellation (PB2 PB1 PA HA NP NA M NS)	HA Clade	Count: Study Samples	Count: EpiFlu Samples
1		1	0	1
2		3A	17	3
3		3B	0	11
4		3A	1	0
5		6	1	2
6		6	0	6
7		5[a]	3	0
8		5	0	3
9		6	1	1
10		5	18	1
11		3C.1, 3C.2, 3C.3, 3C-2012/13	106	16
12		3C.2[b]	7	0

98% nucleotide identity clusters*: Cluster 1, Cluster 2, Cluster 3, Cluster 4, Cluster 5

*Please note that each column is evaluated independently and cluster numbering is arbitrary.

B.

	Genome Constellation (PB2 PB1 PA HA NP NA M NS)	HA Clade	Count: Study Samples	Count: EpiFlu Samples
1		3C.1	1	1
2		3C.1	1	2
3		3C.1	0	1
4		3C.3	1	2
5		3C.3, 3C.2[c]	12	3
6		3C.3[d]	5	1
7		3C.2	20	1
8		3C-2012/13	5	1
9		3C-2012/13	61	0
10		3C.2[b]	7	0

99% nucleotide identity clusters*: Cluster 1, Cluster 2, Cluster 3, Cluster 4, Cluster 5, Cluster 6

*Please note that each column is evaluated independently and cluster numbering is arbitrary.

[a]These study samples form their own HA monophyly within clade 5.

[b]These study samples form their own HA monophyly within clade 3C.2.

[c]Of the 13 strains in this constellation that are also in the HA phylogeny, only one strain falls within clade 3C.2.

[d]These study samples form their own HA monophyly within clade 3C.3.

The authors gratefully acknowledge the 32 originating and submitting laboratories who directly contributed sequences used in the constellation analysis to GISAID: Alaska State Virology Lab; Arizona Department of Health Services; Austin Health, Australia; California Department of Health Services; Canterbury Health Services, New Zealand; US Centers for Disease Control and Prevention; Institute of Medical and Veterinary Science (IMVS), Australia; Institut Pasteur New Caledonia; Iowa State Hygienic Laboratory; John Hunter Hospital, Virology Unit, Clinical Microbiology, Australia; Kentucky Division of Laboratory Services; Melbourne Pathology, Australia; Michigan Department of Community Health; New Mexico Department of Health; New York State Department of Health; Papua New Guinea Institute of Medical Research; Pathwest QE II Medical Centre, Australia; Pennsylvania Department of Health; Puerto Rico Department of Health; Queensland Health Scientific Services, Australia; Research Institute of Tropical Medicine, Philippines; Rhode Island Department of Health; Royal Hobart Hospital, Australia; Southern Nevada Public Health Lab; Spokane Regional Health District, Washington; State of Hawaii Department of Health; Texas Department of State Health Services-Laboratory Services; USAMC-AFRIMS Department of Virology, Cambodia; Utah Department of Health; Victorian Infectious Diseases Reference Laboratory, Australia; WHO Collaborating Centre for Reference and Research on Influenza, Australia; and WHO National Influenza Centre, National Institute of Medical Research (NIMR), United Kingdom.

Figure 27 Human Influenza A H3N2 Genome Constellation Analysis from Huston, Texas During the 2012-2013 Influenza Season. Constellation analysis was conducted at the 98% cutoff for the whole data set in panel (**A**). Panel (**B**) depicts a 99% cutoff constellation analysis conducted on clade 3 viruses only. Reassortments are visible in both panels (4).

The human H3N2 influenza data set included a set of viruses collected during the 2012-2013 North American influenza season in Huston, Texas. In addition to these viruses the analysis also included vaccine viruses as well as manufacturing seeds, a set of viruses used as references in the HA phylogenies of the NIMR September 2012 report and in the CDCs VIRPAC report to the FDA in February of 2012 (97). A sample of sequences from the 2012-2013 season outside Texas was also included as context for the analysis. This data set was examined at the 98 and 99% identity cutoffs (Figure 27). The complete set of genomes was examined at a 98% cutoff, and the predominant circulating clade (3C) (in influenza, phylogenetic clades are defined by the HA gene) and its subclades was examined at 99% for finer resolution of variation. At the 98% cutoff our analysis revealed intrasubtypic reassortment, especially between HA clades 5 and 6. This pattern was confirmed through a comparison of the HA clades and a NA ML tree. Several long-branch-length monophylies were apparent in the NA tree, interleaved with opposite clade NA sequences. Interleaving was also apparent within the 3C clade and subclades. These reassortments were visible in our constellation analysis at a cutoff of 99%. In addition to reassortment, the constellation analysis allows for the visualization of the variability in circulating viruses. During one season, we saw nine distinct viral genotypes at the 98% cutoff level circulating in the Houston area. The ability of influenza to reassort is a key component to its epidemiological success (98).

North American Avian H7 Influenza Diversity Leading to Highly Pathogenic Poultry Outbreaks

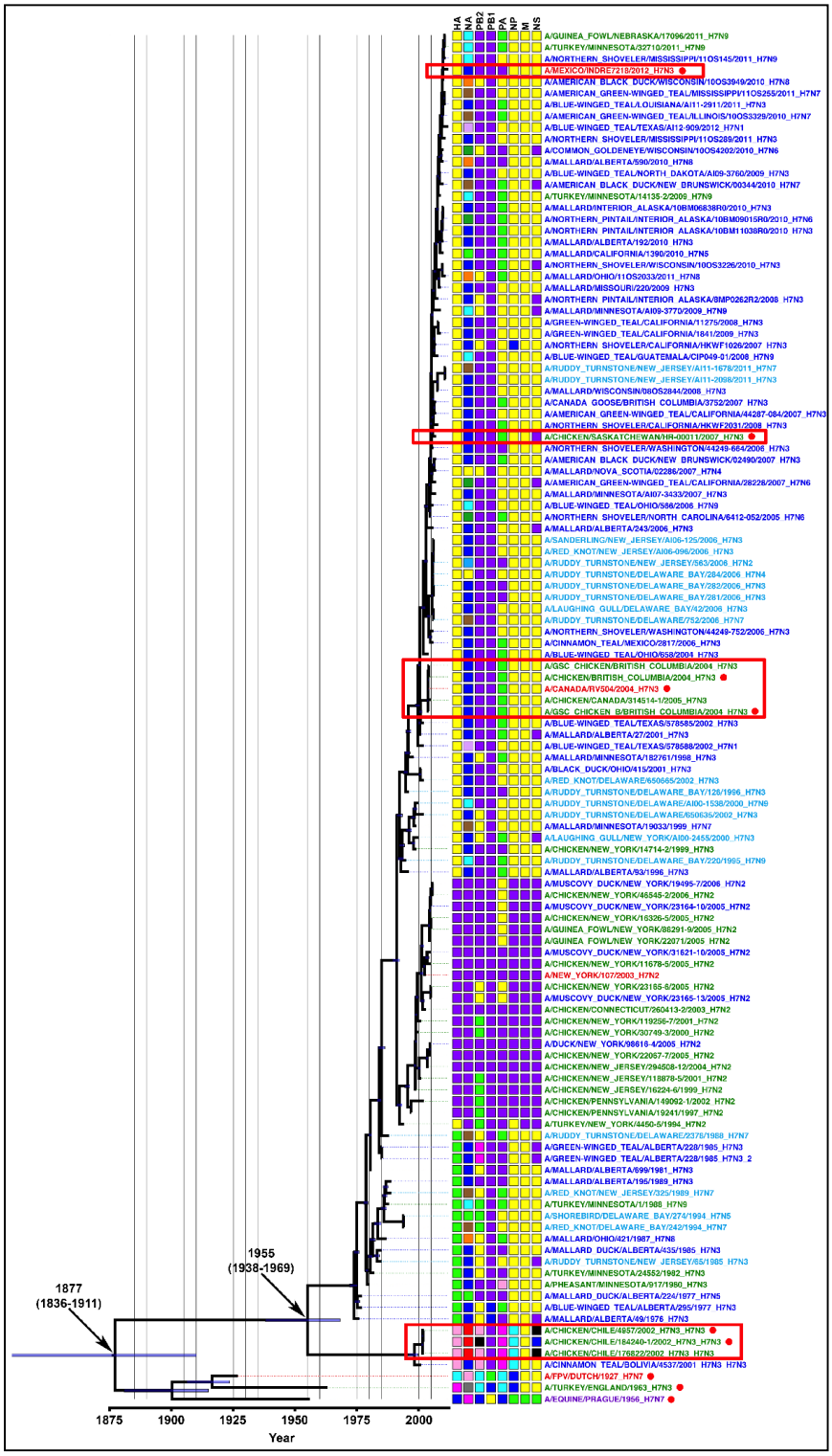


Figure 28 Influenza A H7 Genome Constellation Superimposed on A Bayesian Phylogeny of The HA Gene of Influenza Collected from North American Wild and Domestic Birds. In this figure shore bird, duck, and poultry species are indicated with light blue, dark blue and green strain labels respectively. Divergence date estimates for North and South American H7 lineages are indicated with arrows, as are American and European lineages of H7. H7N3 outbreaks in poultry are indicated with red boxes. Highly pathogenic (HP) H7 outbreaks are indicated with red dots. Each outbreak of HP was revealed to have originated from a different genome constellation (4).

The avian H7 influenza data set included several collections of H7 avian influenza virus genomes from across North and South America, focusing on *Anseriformes*, *Charidraformes* and *Galliformes* (4). In addition to these collections, all available complete genomes from North America were included

for context, as well as several important reference sequences known to have produced highly pathogenic outbreaks of avian H7 influenza in poultry and humans. Although few, if any, complete genomes were available from South America, several nearly complete genomes were included for this analysis.

The constellation analysis was examined at 90 and 95% cluster sequence identity. At both cutoff levels, unique constellations were seen for all highly pathogenic outbreaks (as well as the related low pathogenic poultry viruses) (Figure 28). The wild bird viruses with the HA gene most similar to that of the poultry outbreak viruses were often highly reassorted in comparison to the outbreak viral constellations. This suggests that highly pathogenic outbreaks occur through convergent evolutionary processes (Figure 28 red boxes), and that there are no lineage-specific genes required for this convergent evolution to occur. Although small numbers of surveillance sequences from the time and location of the outbreaks are available, the sequence analysis suggests that these viruses evolve highly pathogenic influenza features once they have infected domestic birds rather than evolving in the wild birds first. Also, *Anseriformes* appear to be the reservoir host for the poultry outbreaks examined here. The constellation diagrams highlight how much missing surveillance data there is about avian influenza as it circulates in wild populations, and how increased avian surveillance will help to fill in these gaps.

Identification of Respiratory Syncytial Virus Chimeras

The RSV data set included a set of viruses collected from children in the US during the 2012-2013 RSV season. In addition to these viral sequences all available complete human RSV genomes were included in this analysis, as the total

complement of full RSV genomes in GenBank is fairly small (as of this writing, approximately five hundred total genomes). After phylogenetic analysis the data set was reduced to 245 genomes including 72 study sequences with complete gene counts.

The RSV genome constellations were examined at 95 and 97.5% cluster sequence identity. At 95% 22 constellations were inferred in the RSV data set (Figure 29). The study sequences fell into five genome constellations: four RSV-A constellations and one RSV-B constellation. All but one constellation with study sequences contained multiple genomes, including non-study genomes. On closer examination of constellation 2, containing just one genome, it was noted that the M2-2 gene more closely matched the M2-2 gene from constellation 14. This pattern continued when examining the genomes at 97.5%; however, the mismatched genomic

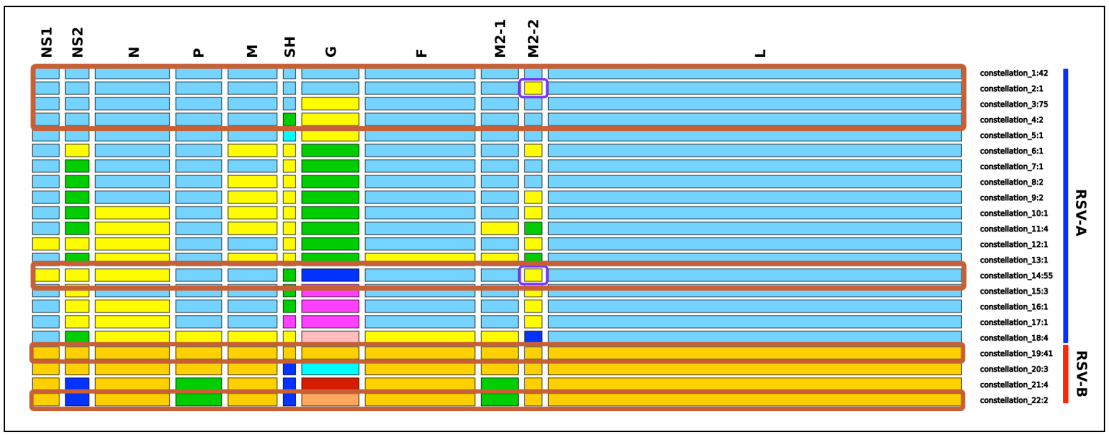


Figure 29 Respiratory Syncytial Virus Genome Constellation Analysis Reveals Putative Recombinant. The genome constellation analysis of RSV highlights the ability of OrionPlot to expand column widths according to markup contained within the matrix tsv file. In this case column widths have been scaled to relative gene lengths. This depiction highlights study sequences with red boxes. A constellation cutoff of 95% was used here. The genome constellations of our study genomes show a potential recombinant depicted with a purple box around the M2-2 gene of constellations 2 and 14.

region expanded to include the M2-1 gene. These results suggest a putative recombination event between the genomes derived from the two separate RSV-A

linages. An examination of the SNPs of this potential recombinant comparing the two parent lineages further supported a recombination event, as did the results from using the RDP3 recombination detection software. Taken together these results strongly support the hypothesis that recombination occurred between these two RSV-A lineages to form the virus in constellation 2.

To confirm the recombination hypothesis, the sequencing methodology for this genome project was reviewed. The study viruses were sequenced by first making four PCR amplicons tiled across the genome. The PCR amplicons are then pooled and sequenced using the IonTorrent platform. The PCR primers used for the amplification were compared to the coordinates of the putative recombination event. The coordinates for the primers for amplicon 3 were within 100 nucleotides outside the location of the putative recombination site. As part of the assembly process, a deep sequence analysis of the assembly was produced. Variant sites were detected in the assembly within these same regions where ampicon 3 overlapped with amplicon 2 and amplicon 4. These results suggest the putative recombinant was derived artificially as a result of an error in the sequencing methodology, specifically at the amplicon pooling step. To confirm this, PCR primers were designed flanking the putative recombination site. PCR products were obtained for the amplicon 3 and 4 junction and were Sanger sequenced. The resultant sequences were aligned to the major parent, the minor parent and the putative recombinant using MAFFT (73) and visualized using CLCbio (96). All PCR product derived sequences matched the major parent, demonstrating the likelihood of a methodologically derived error in the sequencing process resulted in a false positive detection of recombination.

Diversity in Zaire Ebola Virus and Identification of Protein Variants

The Ebola virus data set includes all complete Zaire Ebola virus genomes available as of September 2014. The majority of the genomes available were from the recent Ebola virus outbreak in West Africa (99). Ebola virus has a relatively low mutation rate compared to many RNA viruses (100). Few non-synonymous mutations were observed in the majority of the Ebola genes during the early outbreak sampling. The goal with this analysis

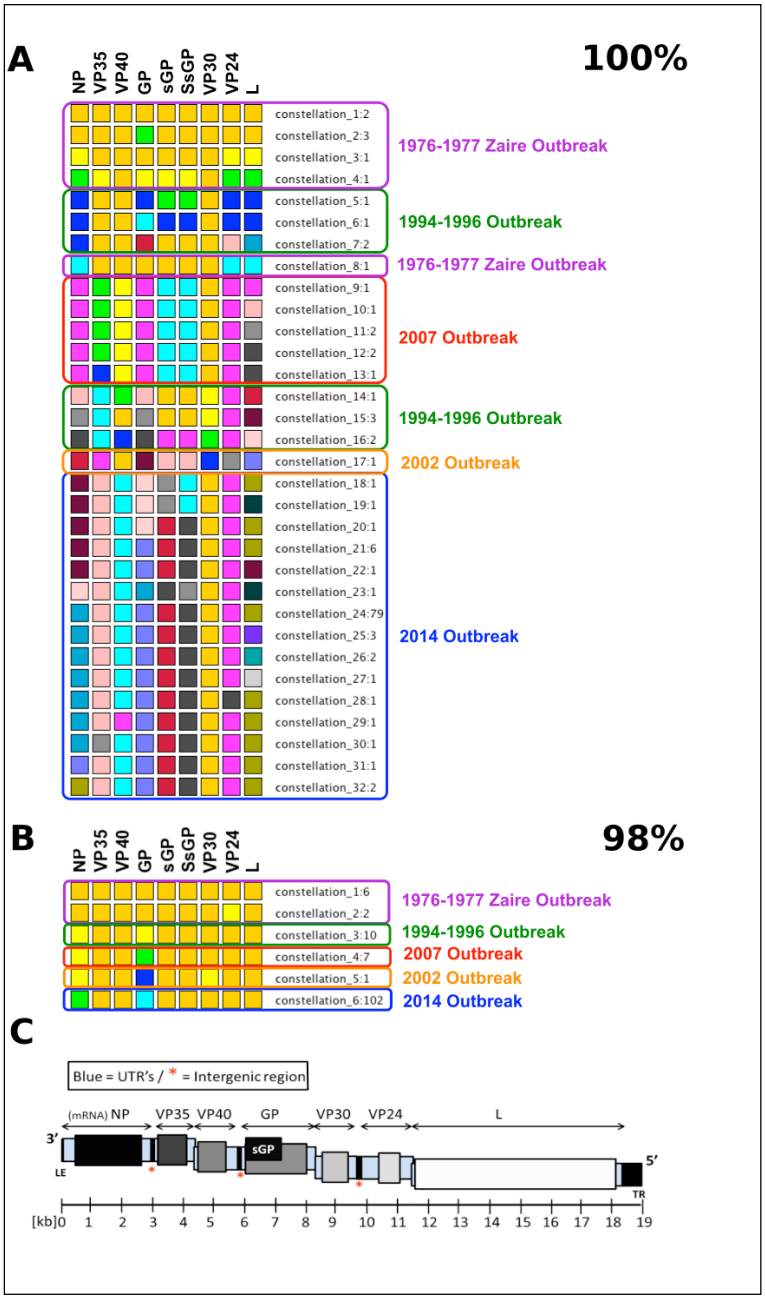


Figure 30 Zaire Ebola Virus Protein Constellation Plot. (**A**) Protein constellation at a 100% cutoff. Variability in the each gene is apparent. (**B**) At the 98% cutoff no within outbreak constellation differences are present for all known Zaire Ebola outbreaks. It is worthy to note how little variation occurs in Ebola within and between outbreaks. The Zaire outbreak genomes include several of the same strain that have been resequenced with variable results. (**C**) The genome structure has been included for reference. Thanks to Reed Shabman for providing this genome diagram.

was to screen the variant viruses that had some level of detectable mutation, so that assays could be performed to assess the effect of the mutations on the viral protein phenotypes *in vitro*. The available Ebola virus genomes were annotated with VIGOR 3.0 (22) and the protein sequences were binned by gene product into amino acid fasta files. Here we used the amino acid sequence since we were primarily interested in screening protein variants for further study. We set the cutoff at 100% identity, meaning sequences with only one amino acid change would establish a new protein cluster and thus new constellation. The results of this clustering revealed just 32 distinct Ebola virus constellations in all the known Zaire Ebola virus genomes available in GenBank at that time (September 2014) (Figure 30). Twenty L protein, fifteen NP protein, fourteen GP protein, ten sGP protein, nine SsGP protein, nine VP24 protein, four VP30 protein, eight VP35 protein, and six VP40 protein variants defined these constellations. The 2014 outbreak showed 15 distinct constellations, with more than two variants observed in just five of the nine Ebola virus proteins we used to generate the constellations. Of the proteins with mutations, three variants were uncovered in the GP, sGP, and SsGP proteins, five in the NP protein, and six in the L protein. This low level of variation is consistent with previous studies on Ebola virus evolutionary rates (99, 100).

Comparison of Constellation Clusters to Phylogenies

A comparison of constellation clusters to phylogenetic trees generated using the maximum likelihood estimation method revealed cluster dendrograms with similar overall topologies to the respective phylogenetic trees – with a couple notable exceptions. First, as seen in the RSV example, percent identity distances derived

from indels in the sequence alignments were calculated differently in the VirComp pipeline as compared to phylogenetic methods. This can been seen when comparing the phylogenetic placement (Figure 33, Chapter 4) of the four RSV strains (LA2_27, LA2_39, LA2_45, and LA2_78) in the lower group (TN1) of viruses to their placement in the constellation clustering (Figure 29) (where they are placed with the top group of viruses, constellation 1). A closer examination shows that these viruses have a variation pattern more similar to the group of viruses they are placed with in the phylogeny. A closer examination of the alignments shows a large indel of 72 nucleotides within the G protein. The G protein is the only gene that has a different cluster assignment between constellations 1 and 3. In constellation 3 the insertion does not appear, whereas constellation 1 the insert does appear. The strains that matched the lower phylogenetic group (but are placed with constellation 1 strains) have the insert. It would appear that the added distance calculated from the mismatch (derived from gap to insertion sequence) is enough to group these strains in constellation 1, even though evolutionarily they appear to be closer relatives to the constellation 3 viruses.

Second, at times, members of constellation clusters on the edge of phylogenetic clades sometimes appeared to have questionable cluster assignments. This assignment disparity may be due to the differences in the way that the phylogenetic methods evaluate evolutionary distances compared to the strict accounting of percent identity used by the VirComp software.

Discussion

Detection of Genomic Rearrangements

The VirComp method for constellation analysis of viral genomes provides a useful framework for generating publication worthy visualizations of genome constellations. These visualizations allow users to establish evidence of reassortment in segmented viruses and can also be used to demonstrate recombination in non-segmented genomes. By creating visualizations at several cutoff levels, the relative age of these reassortment events can be established as well. Although it may be important to verify these findings with one of the more rigorous reassortment detectors, such as GiRaF (101), VirComp provides a fast and context-specific approach to constellation analysis. As demonstrated from the RSV example, the detection of genomic rearrangements is not limited to naturally derived rearrangements. As such, this method could be coupled with the genome assembly process to detect errors in sequence production and to confirm (or rule out) true genome rearrangements (Figure 31).

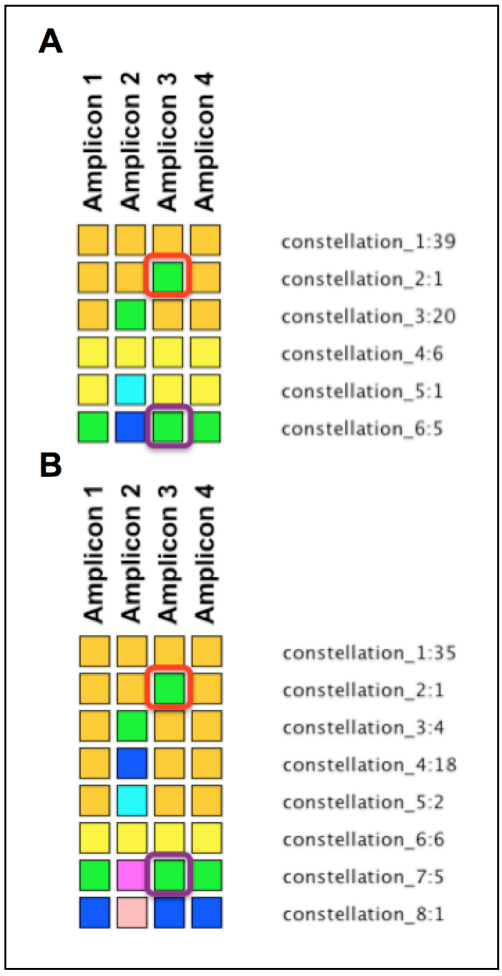


Figure 31 Reanalysis of Putative RSV Recombinant Virus in the Context of Study Sequences Only and Lab-generated Amplicon Segments. (**A**) Study samples examined by amplicon at 98% cutoff (**B**) The same data set at the 99% cutoff. These plots show RSV recombinant over laps with the region of the third amplicon we used to amplify and sequence the genomes. This finding suggests in the constellation with the amplicon in the purple box was mispooled into the genome with the red box prior to sequencing.

Species Agnostic

VirComp has been demonstrated to be a species agnostic method, in that any viral species gene complement can be compared with this methodology and the code base is not specific to either influenza or RSV. This fact suggests that comparisons in other domains of life could be performed, such as analyses on the pathogenicity of specific genes across several bacterial strains. RNA virus genomes especially lend themselves to this type of comparative analysis, due to their small genome size.

Diversity and Visualization

Biological relevance and ease of visualization can be balanced with the context specific diversity in the data sets being analyzed by selecting a meaningful cutoff. Furthermore, cutoffs can be adjusted recursively to look more closely at specific subsets of the diversity being surveyed, as evidenced by the analysis of H3N2 viruses described herein. This allows for the examination of data sets at various levels of diversity, and of the timing of specific diversifying events (such as genomic recombination and reassortment).

Variant Selection

As evidenced by the Ebola example, with strict cutoffs (100%) in amino acid space, protein sequence variants can be quickly screened and analyzed for further study in lab-based assays.

Performance and Algorithms

VirComp uses fast, deterministic algorithms for its comparisons. Farthest neighbor is a deterministic distance-based hierarchical clustering algorithm that will

produce only one result for a given input data set (102). This allows for reproducibility. These types of hierarchical clustering algorithms are fast, especially then compared to maximum likelihood or Bayesian phylogenetic approaches (which are also non-deterministic). An analysis of several hundred viruses with eight genes may take on the order of 30 minutes with VirComp. A similar Bayesian or maximum likelihood analysis could take hours or days depending on the complexity of the evolutionary model used and the size of the data set. Furthermore, farthest neighbor coupled with the use of Hamming distances as implemented by the ANDES software (39) provides a distance metric (percent identity) which is easy to measure and therefore provides a consistent mode for establishing clusters that is more readily understandable across analyses.

Drawbacks and Limitations

It should be noted that there are a couple of drawbacks to the approach VirComp uses. First even though simple distance-based calculations are fast, they do not always provide the best reconstruction for the evolutionary history of the sequences being examined. This fact could lead to localized miscategorizations of genes made by using sequence similarities that could have derived from convergent evolutionary events. Our analysis suggests the main differences seen between the VirComp and the phylogenetic-based methods are seen at the boundaries of clades. A second drawback is that the distance method used is incapable of calculating distance differently when a difference between two genomes is derived from a single evolutionary event or several events. An example of this was seen with in our RSV data set. Four genomes in constellation 1 should have been assigned to constellation

3. Our method calculated more distance between these four viruses and their correct constellation 3 because of a 72bp insertion in the G gene of the four genomes in question. This insertion matched one seen in all the genomes of constellation 1. This similarity in duplication status was enough to miscategorize the four genomes with constellation 1, despite having more similarity to constellation 3 at all other sites within G. Essentially, the distance only approach treats each additional base as a separate evolutionary event, and thus calculates more distance than would be inferred using an evolutionary model that can score indels appropriately, or simply treat them as missing data.

Similar Methods

The literature highlights a few similar methods to this type of constellation analysis. An example of this is a heat map style approach to compare a single reference strain to a small collection of viruses of interest, such as presented in *Gonzalez-Reiche et al* (50). This type of approach is only useful when the context of the question of similarity is 2-dimensional, i.e. a single subject compared to a small set of references. This type of comparison establishes a one-to-many relationship centered on the reference in question. This methodology does not provide comparisons of the various comparators, and thus these analyses are prone to miss important relationships between those comparators.

Another method highlighted in the literature is that of RotaC (91) and FluGenome (51). These methods, although on a more firm methodological footing, in that they do the necessary all-versus-all comparison of genome components, they do require a more rigorous computation and maintenance of the genomic

constellation categories inferred over time. Although this is highly useful for putting a particular set of virus constellations into the context of all the known viral strains of a particular species, this type of analysis will become increasingly complex as the databases of viral sequences expand. It also requires the adherence to an established nomenclature for genotypes that may be difficult to establish, especially in large viral research communities. VirComp mitigates that issue simply by providing only a study specific analysis and does not aim to integrate its constellation results with established or yet to be established nomenclatures. Once such nomenclatures have been established it should be a goal to extend those standards to the constellation annotation features of VirComp. Additionally, the RotaC approach applies differing cutoffs for the various genes. This makes it difficult understand quantitatively how much mutation has occurred between defined genotypes.

Conclusions

In summary, here we describe a novel constellation analysis method and visualization software useful in performing viral comparative genomics analyses. This method is a distance-based approach that utilizes several preselected cutoffs to create genome constellations at various levels of resolution. Although, this method does not perform phylogenetic reconstructions (and therefore does not provide evolutionary clade-based genotypes), it does employ a straightforward percent identity metric that is systematic and not arbitrary. We have also provided five examples of using this method to perform viral surveillance, reassortment analysis, protein variant screening, and recombination and sequencing chimera screening. The ease of use and fast runtime of VirComp and OrionPlot make this methodology an

important advancement in viral comparative genomics – especially in light of the certain expansion of viral genomics using next-generation sequencing technologies.

Chapter 4: Large-Scale Respiratory Syncytial Virus Whole-Genome Sequencing Identifies Sequence Duplication in G Gene Associated with Reduced Diseases Severity

Abstract

Respiratory syncytial virus (RSV) is the most important respiratory pathogen for children under the age of five. RSV is responsible for considerable morbidity and mortality worldwide. Variability within and between RSV group A and B viruses and the ability of multiple strains of RSV to co-circulate are likely mechanisms for the evasion of herd immunity. Detailed studies following the whole-genome variability of RSV over time are required to understand these dynamics and to understand the phenotypes these variants possess. In this study, we performed complete-genome next-generation sequencing of 71 RSV isolates from infants in central Tennessee during the 2012-2013 RSV season. We identified multiple co-circulating clades of RSV from both the A and B groups. Each clade is defined by signature N- and O-linked glycosylation patterns. A detailed look at specific RSV genes revealed high rates of positive selection in the attachment (G) gene. Furthermore, we identified RSV-A viruses in circulation with and without the recent 72-nucleotide G gene duplication. In our study, this duplication appears to be associated with less severe RSV infections. Pairing of high-throughput next-generation sequencing with a large cohort study of RSV-infected infants enabled a detailed look at both molecular evolutionary dynamics and exploratory clinical phenotype analyses looking for sequence-based signatures of disease.

Introduction

Respiratory syncytial virus (RSV) was first isolated in 1955 (83, 103) and has been associated with mild to severe acute lower respiratory infections (ALRIs), especially in infants, premature babies, the elderly and immunocompromised individuals (85, 104-106). In 2005, RSV caused an estimated 33.8 million new episodes of ALRIs in children under five worldwide, with 3.4 million cases requiring hospitalization due to severe illness (83, 85, 86). Global estimates of disease burden show RSV to account for 30 million ALRIs and 50,000 annual deaths of children < five years of age (83, 85, 86). Nearly all children have had at least one RSV infection by two years of age (85). However, a recent study has established that RSV infections during infancy (less than six months of age) are associated with an increased incidence of subsequent childhood wheezing and asthma (43). Despite its global public health impact, no licensed vaccines or effective medications for young children infected with RSV are currently available (43). The only approved prophylaxis is passive immunization with palivizumab (Synagis from MedImmune), a humanized mouse antibody (mAb) against the RSV F protein (107, 108). The original clinical trial of palivizumab indicated a 39-78% decrease in hospitalization rates for RSV in premature infants and children with chronic lung disease, although subsequent analyses have suggested lower affectivity (107).

RSV is an enveloped virus with a negative-sense, single-stranded, non-segmented RNA genome belonging to the Paramyxoviridae family. The 11 RSV proteins are: the polymerase (L) protein, the nucleocapsid (N) protein, the phosphoprotein (P), the transcriptional regulators (M2-1 and M2-2) proteins, matrix

(M) protein, the small hydrophobic (SH) protein, the non-structural proteins (NS1 and NS2) proteins, and two major surface glycoproteins (F and G). The F and G proteins are responsible for virus entry and are the major target of human immune responses. The F protein is responsible for the fusion of the viral envelop with the host cell membrane. The G protein is responsible for cellular attachment and has an immune decoy function in its soluble, extracellular, secreted form. The G protein is organized from N to C terminal as follows: cytoplasmic domain, transmembrane domain, mucin-like domain, conserved domain, attachment domain, and finally a second mucin-like domain.

RSV has an epidemic seasonality, with increased cases during the winter in temperate climates and during the monsoon season in tropical climates (43, 83, 109). RSV can be classified into two antigenic groups (A and B), each containing several distinct subgroups based on antigenic and genomic sequence differences, especially in the G glycoprotein (110, 111). Studies suggest group A viruses cause more severe disease and transmit more readily than group B viruses in infants (43). These two groups tend to alternate in prevalence between RSV seasons. There is also evidence of multiple co-circulating intra-group viral genotypes, or clades, during any given season (109, 111), resulting in a diverse set of circulating viruses that can adapt to herd immunity. It is unclear if this represents a gradual evolution of the viral genomes or stochastic differences in infection rates by co-circulating strains.

Previous RSV sequencing studies have largely focused on sequencing complete or partial G gene sequences because the C-terminal (the second hypervariable portion of G) is sufficient and required for distinguishing the two RSV

groups and the various genotypes within each group (109, 112). In 1999, a G gene variant was identified in RSV-B that contained a 60-nucleotide (20 amino acid) duplication in the C-terminal third of the G gene, within the second mucin-like domain (113, 114). This genotype has now spread globally (115). In 2011, a similar G gene variant was identified in RSV-A from several locations around the globe that contained a 72-nucleotide (24 amino acid) duplication in the second mucin-like domain (110, 115, 116).

To better understand RSV evolutionary dynamics, we sequenced RSV genomes from acutely infected infants from middle Tennessee who were enrolled as part of the Infant Susceptibility to Pulmonary Infections and Asthma Following RSV Exposure (INSPIRE) study. The objective of our sequencing efforts was to identify new variants in these RSV genomes and investigate any correlations with clinical aspects of RSV-associated upper respiratory infections (URIs).

Methods

Study Population

INSPIRE is an observational, population-based, longitudinal study of previously healthy, full-term infants enrolled near birth, which conducts surveillance for respiratory illnesses during their first winter viral season. Eligible infants were born between June and December and were on average 6 months of age or older during sampling for this study. Informed consent was obtained from the legal guardians of each infant. All procedures were in accordance with the ethical standards of the Vanderbilt University Institutional Review Board. Demographic

data – including age, sex, race and ethnicity – were recorded at the time of enrollment.

RNA Extraction and RT-PCR

Extraction of the viral RNA was performed at the J. Craig Venter Institute (JCVI) in Rockville, MD with 140 µl of nasal wash sample using the ZR 96 Viral RNA kit (Zymo Research Corporation, Irvine, CA, USA). Four forward reverse transcription (RT) primers were designed and four sets of PCR primers were manually picked from primers designed across a consensus of complete RSV genome sequences using JCVI's automated primer design tool (12). The four forward reverse transcription (RT) primers were diluted to 2 µM and pooled in equal volumes. cDNA was generated from 4 µl undiluted RNA, using the pooled forward primers and SuperScript III Reverse Transcriptase (Thermo Fisher Scientific, Waltham, MA, USA). Four independent PCR reactions were performed on 2 µl of cDNA template, using either AccuPrime Taq DNA Polymerase (Thermo Fisher Scientific) or Phusion High Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA, USA) to generate four overlapping amplicons (approximately 4-kb each) across the genome. Amplicons were verified on 1% agarose gels, and excess primers and dNTPs were removed by treatment with Exonuclease I (New England Biolabs) and shrimp alkaline phosphatase (Affymetrix, Santa Clara, CA, USA) for 37°C for 60 min, followed by incubation at 72°C for 15 min. Amplicons were quantitated using a SYBR Green dsDNA detection assay (SYBR Green I Nucleic Acid Gel Stain, Thermo Fisher Scientific), and all four amplicons per genome were pooled in equal volumes.

RSV Whole-Genome Sequencing

For samples sequenced using the Ion Torrent PGM (Thermo Fisher Scientific), 100 ng of pooled DNA amplicons were sheared for 7 min, and Ion-Torrent-compatible barcoded adapters were ligated to the sheared DNA using the Ion Xpress Plus Fragment Library Kit (Thermo Fisher Scientific) to create 400-bp libraries. Libraries were pooled in equal volumes and cleaned with Ampure XP reagent (Beckman Coulter, Inc., Brea, CA, USA). Quantitative PCR was performed on the pooled, barcoded libraries to assess the quality of the pool and to determine the template dilution factor for the emulsion PCR. The pool was diluted appropriately and amplified on Ion Sphere Particles (ISPs) during emulsion PCR on the Ion One Touch 2 instrument (Thermo Fisher Scientific). The emulsion was broken, and the pool was cleaned and enriched for template-positive ISPs on the Ion One Touch ES instrument (Thermo Fisher Scientific). Sequencing was performed on the Ion Torrent PGM using 316v2 or 318v2 chips (Thermo Fisher Scientific).

For samples requiring extra coverage, in addition to Ion Torrent sequencing, Illumina libraries were prepared using the Nextera DNA Sample Preparation Kit (Illumina, Inc., San Diego, CA, USA) with half reaction volumes. Briefly, 25 ng of pooled DNA amplicons were tagmented at 55°C for 5 min. Tagmented DNA was cleaned with the ZR-96 DNA Clean & Concentrator Kit (Zymo Research Corporation) and eluted with 25 µl of resuspension buffer. Illumina sequencing adapters and barcodes were added to tagmented DNA via PCR amplification, where 20 µl of tagmented DNA was combined with 7.5 µl of Nextera PCR Master Mix, 2.5 µl of Nextera PCR Primer Cocktail and 2.5 µl of each of index primer (Integrated

DNA Technologies, Coralville, IA, USA) for a total volume of 35 µl per reaction. Thermocycling was performed with 5 cycles of PCR, as per the Nextera DNA Sample Preparation Kit protocol (3 min at 72°C, denaturation for 10 sec at 98°C, annealing for 30 sec at 63°C and extension for 3 min at 72°C) to create a dual-indexed library for each sample. After PCR amplification, 10 µl of each library was pooled into a 1.5 mL tube, and the pool was cleaned twice with Ampure XP reagent (Beckman Coulter, Inc.) to remove all leftover primers and small DNA fragments. The first cleaning used a 1.2x volume of the Ampure reagent, whereas the second cleaning used a 0.6x volume of the Ampure reagent. The cleaned pool was sequenced on the Illumina MiSeq v2 instrument (Illumina, Inc.) with 300-bp paired-end reads.

RSV Genome Assembly and Annotation

Sequence reads were sorted by barcode, trimmed, and de novo assembled using CLC Bio's *clc_novo_assemble* program (16). The resulting contigs were searched against custom, full-length RSV nucleotide databases to find the closest reference sequence. All sequence reads were then mapped to the selected reference RSV sequence using CLC Bio's *clc_ref_assemble_long* program (18). At loci where both Ion Torrent and Illumina sequence data agreed on a variation (compared with the reference sequence), the reference sequence was updated to reflect the difference. A final mapping of all next-generation sequences to the updated reference sequences was performed with CLC Bio's *clc_ref_assemble_long* program (18). Curated assemblies were validated and annotated with the viral annotation software called Viral Genome ORF Reader, (VIGOR) 3.0 (22), before submission to GenBank. VIGOR was used to predict genes, perform alignments, ensure the fidelity of open

reading frames, correlate nucleotide polymorphisms with amino acid changes, and detect any potential sequencing errors. The annotation was subjected to manual inspection and quality control before submission to GenBank. All sequences generated as part of this study were submitted to GenBank as part of the Bioproject ID PRJNA225816.

Phylogenetic Analyses

Sequence collection **–** All available full-length human RSV-A and RSV-B genomes were downloaded from GenBank on June 24, 2014. Any viral isolates that contained "mutant" or other key words indicating *in vitro* modifications were removed from the data set after an initial ML phylogenetic analysis. The remaining public genomes were then combined with the 71 RSV genomes from the study samples collected during the 2012-2013 winter RSV season. Full genomes were then annotated using VIGOR 3.0 (22) to ensure consistent gene annotations across all genomes. Each of the 11 RSV genes were separated into gene-specific fasta files for gene-based phylogenetic analyses. Although in principal this data set should have included only complete gene sets, genomes without complete gene counts or containing partial gene annotations after processing with VIGOR were excluded from further analysis.

Maximum likelihood analyses of whole-genome and G-gene-specific RSV sequences **–** The nucleotide substitution model used for all phylogenetic analyses was a general time reversible model with a nucleotide site-specific rate heterogeneity with four rate categories, and invariant sites (GTR-IG, as determined by jModelTest2.4) (117). MAFFT (93) was used to create whole-genome and G-gene-specific alignments, and all alignments were checked and edited as appropriate. Maximum

likelihood phylogenies were inferred using an adaptive best tree search on the GARLI Web Service 2.0 (95) to statistically ensure the best tree (as measured by log likelihood scores) was found over 1000 replicates. The resultant tree was labeled with the viral strain names, and colored using in-house PERL scripts. Clade designations for the study sequences were defined using the reduced G gene phylogeny (S2) by examining boostrap support on branches leading to clades with study sequences. All branches leading to the study sequences were supported by boostrap values > 70.

Bayesian phylogenetic analyses of RSV-A and RSV-B genomes – The whole-genome maximum likelihood tree was used as a guide to select a subset of viral genomes for Bayesian phylogenetic analyses, including genomes with unique phylogenetic histories and commonly used reference genomes. To determine if our data exhibited temporal qualities, we performed an exploratory analysis with Path-O-Gen (available at http://tree.bio.ed.ac.uk/software/pathogen/). Neighbor joining trees generated with RSV-A-only and RSV-B-only genomes were used to measure root-to-tip divergence using Path-O-Gen, which showed that both RSV data sets contained enough temporal signal to proceed with time-based Bayesian analyses. All Bayesian analyses were performed using BEAST v1.8 (38, 118) on the CIPRES Science Gateway (33). Whole-genome and gene-specific phylogenies were inferred using Markov chain Monte Carlo sampling chains (100 million to one billion in length), with parameters and trees recorded to ensure 10,000 samples per run. The GTR-IG substitution model was used and tip dating with precision to the sampling year was employed for all trees. All genes were analyzed using a lognormal relaxed clock.

We utilized the SkyGrid (119) coalescent model, with 50 partitions over 50 years. Default priors were used for each analysis – except for the ucld.mean prior for which we used the CTMC rate reference prior (120). For the G gene analysis, we performed divergence dating on RSV-A by constraining the four clades and genotypes most closely related to two separate lineages of viruses with G-gene duplications present. All analyses were evaluated with Tracer v1.6 (available at http://tree.bio.ed.ac.uk/software/tracer/) to determine the success of the chain sampling based on effective sample size values for each parameter, and additional chains were run as needed. For each analysis, we constructed a maximum clade credibility tree using TreeAnnotator v1.8.1, available for download with BEAST.

Glycosylation Prediction

Two surface glycoproteins, F and G, were analyzed using NetNGlyc (121) and NetOGlyc (122) software. Multifasta files were loaded into the web interface and the output was saved, and then parsed with custom PERL scripts to produce a spreadsheet of glycosylation sequons and amino acid coordinates for the N-linked glycans and amino acid coordinates for the O-linked glycans. The coordinates were then used to produce visualizations in R, using the ggplot2 package (123), of the overall consensus glycosylation patterns on the G protein for the various clades and genotypes identified in this study.

Statistical Analyses

The distribution of categorical clinical characteristics was compared across RSV sequence characteristics using the chi-squared test, whereas continuous

covariables were compared using the Wilcoxon rank-sum test. The BSS2 metric was used to compare against genetic data, since several study samples were missing values using the BSS metric, which requires clinicians to use a pulse oximeter during site visits. All statistical tests were performed in R (124), and visualizations were done with the R package, ggplot2 (123).

BaTS Analysis for Detecting Global Versus Local Circulation Patterns with Tree Topologies

The Bayesian RSV-A G gene phylogeny data set was used to analyze signals of global versus local circulation within our data set. Local versus global circulation status was assigned to each G gene sequence. BaTS analysis (58) was performed using a total of 9001 Bayesian phylogenies of RSV-A G gene sequences from the previous analysis. The BaTS analysis was run in single mode with 100 replicates using two states (local and global).

Results

Large-Scale RSV Whole-Genome Sequencing from Nashville, Tennessee During the 2012-2013 Season

The INSPIRE study was designed to collect samples from study participants to identify any respiratory viral infections during the RSV season in an unbiased manner. A selection of nasal wash samples from patients with acute respiratory tract infections was screened for RSV using qRT-PCR. One hundred six samples from 99 patients were selected for whole-genome sequencing based on the sampling season

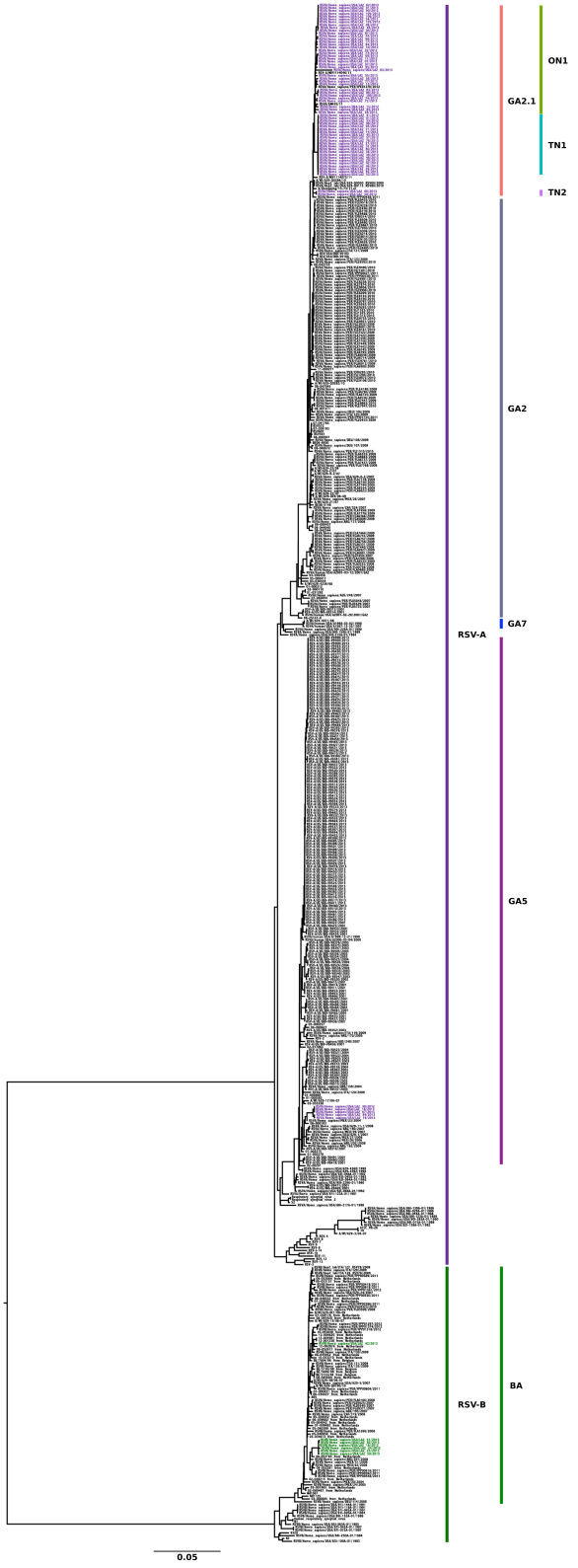
and positive RSV serology results. Characteristics of these study subjects are listed in Table 9.

Table 9 Demographics and Clinical Characteristics of Enrolled Infants (n=99)

Demographics and Clinical Characteristics	Infants with RSV ARTIs (n=99)*†
Age (months)	2.97 (1-7)
Female	42 (42.42%)
Male	57 (57.57%)
Race	
Black	15 (15.15%)
White	72 (72.73%)
Other§	12 (12.12%)
Hispanic ethnicity	11 (11.11%)
Gestational age (weeks)	39 (39-40)

ARTIs = acute respiratory tract infections; RSV = Respiratory Syncytial Virus
*Data are presented as the number (%) for binary variables or median (interquartile range) for continuous variables.
†Percentage calculated for children with complete data.
§Category includes subjects of mixed race.

Six patients were shown to be RSV-positive twice and one patient was RSV-positive three times during the 2012-2013 season. However, in only four patients were we able to obtain consensus sequences. For one of those infants, sequence data was obtained for both samples and the consensus sequences were identical. The clinical data from the second isolation case was excluded from the statistical analysis, but was included in the phylogenetic analyses. Infants were evenly enrolled in the study by sex; most were non-Hispanic, white, and were between one and seven months old at the time of sample collection, with a mean age of 2.97 months (standard deviation = 1.62).



Of the 106 RSV-positive study samples, 71 whole-genome sequences were obtained, annotated and submitted to GenBank. Partial genome sequences were obtained for three additional samples that contained gaps, lower coverage areas, or otherwise did not meet quality standards, because these incomplete genome sequences prevented accurate gene-specific analysis, they were removed from the data set.

Maximum Likelihood Phylogenetic Analyses Demonstrate the Convergent Emergence of G Duplications

Figure 32 Maximum Likelihood Phylogeny of 545 RSV Whole-Genome Sequences Including 474 Downloaded from GenBank on June 24 2014 and 71 Study Sequences. Whole-genomes sequences include mutant and lab-constructed strains. RSV-A study isolates are depicted in purple, whereas RSV-B study isolates are depicted in green. Co-circulating strains from multiple clades were present during the 2012-2013 RSV season in central Tennessee.

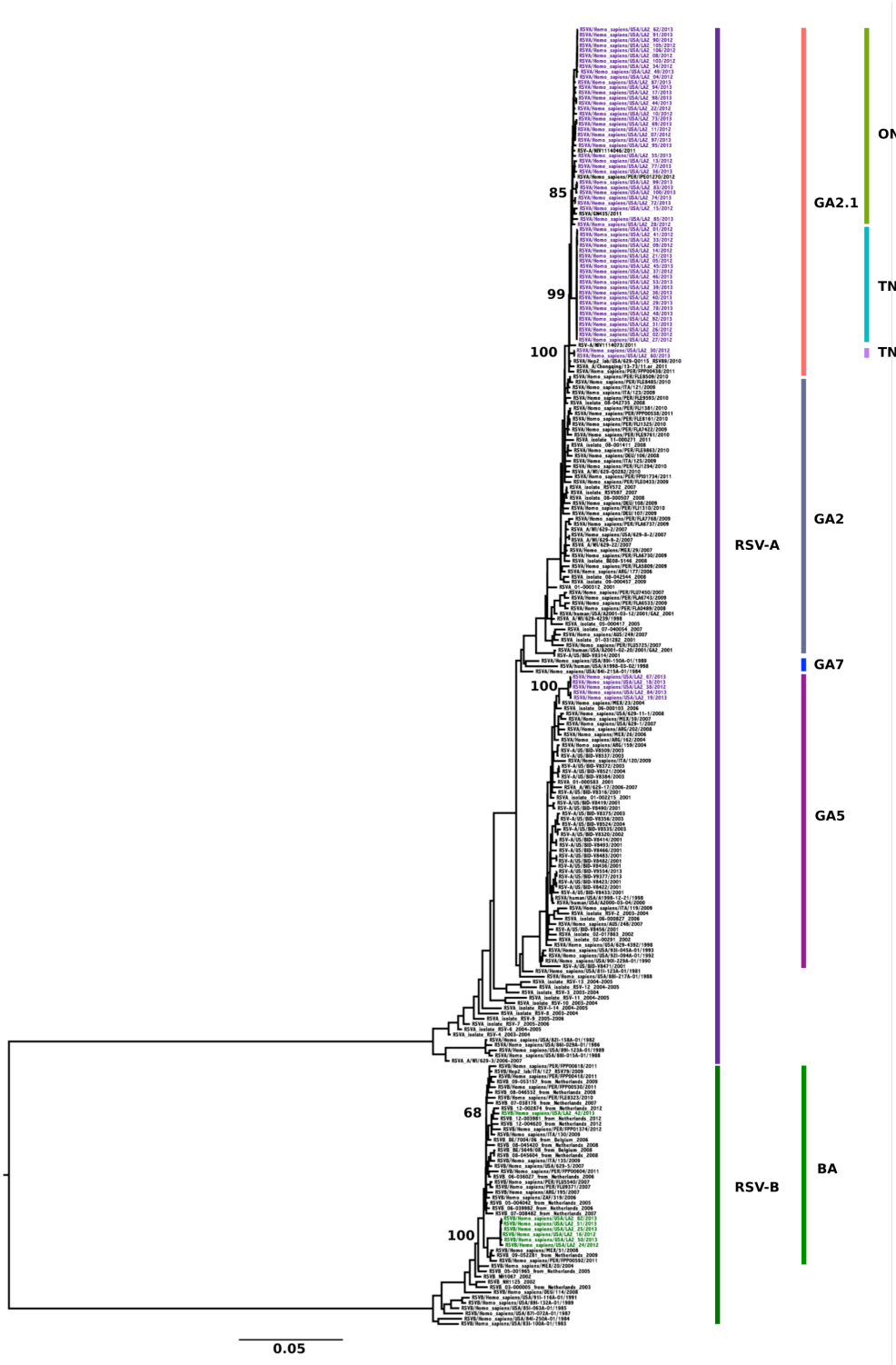


Figure 33 Maximum Likelihood Phylogeny of RSV G Gene Sequences from a Pruned Whole-Genome Data Set. Lab mutants, redundantly sequenced strains, and over-represented branches in the whole-genome phylogeny were removed or thinned out for subsequent analysis of the G gene coding region. RSV-A study isolates are depicted in purple, whereas RSV-B study isolates are depicted in green. Bootstrap support was included for nodes that were important for establishing the clades and genotypes in the Bayesian divergence dating analysis. Bootstrap support on these nodes was > 80% in all cases.

A maximum likelihood phylogeny that combined the RSV-A and RSV-B lineages was generated using whole-genome sequences from 474 publicly available

sequences and the 71 study genomes from the 2012-2013 season (Figure 32). A subset of these samples was used to infer a maximum likelihood phylogeny using only the G gene coding sequence, and a similar topology was obtained (Figure 33). The 71 study genomes were aligned to three separate clades: BA RSV-B (seven genomes), GA5 RSV-A (five genomes) and GA2 RSV-A (59 genomes). The RSV-A clade GA2.1 (a continuation of the GA2 clade) recent isolates were further divided into three monophilies, representing genotype ON1 with 35 genomes. These genomes are a new group of viruses, specific to the study samples from Tennessee, that we are calling genotype TN1. The TN1 genotype has 22 sequences, and has two genome sequences proximal to the divergence point of GA2.1 that we have named genotype TN2 for this study. This confirmed the co-circulation of multiple RSV clades and genotypes in one season and in the same geographical location. Interestingly, seven RSV-B and 39 RSV-A study sample genomes contained a previously reported (110, 113, 115) insertion within the C-terminal third of the G gene coding sequence. The insertion is present as an exact, tandem, in-frame duplication of the same gene region in both the RSV-B and RSV-A genomes, but is 60 nucleotides in length in RSV-B and 72 nucleotides in RSV-A.

Phylogenetic and sequence analyses of the G gene duplications suggest that the duplication occurred convergently at separate times in the RSV-A GA2.1 genotypes (ON1 and TN1), as well as in the RSV-B group (Figure 32). All 35 strains from genotype ON1 contained the G gene duplication, whereas only four out of 22 TN1 genomes contained the G gene duplication. All seven RSV-B genomes

contained the G gene duplication, whereas none of the RSV-A GA5 genomes had the G gene duplication.

Bayesian Phylogenetic Analysis Provides Estimates of RSV Evolutionary Dynamics

Maximum clade credibility (MCC) trees were constructed using the G gene analyses for both RSV-A and RSV-B (Figure 34). Bayesian analyses of each individual gene, as well as the whole genome, provided mutation rates similar to those reported in previous studies (43, 106) for all RSV genes (Figure 35). We also observed a high rate and Bayesian highest posterior density interval (HPD) of

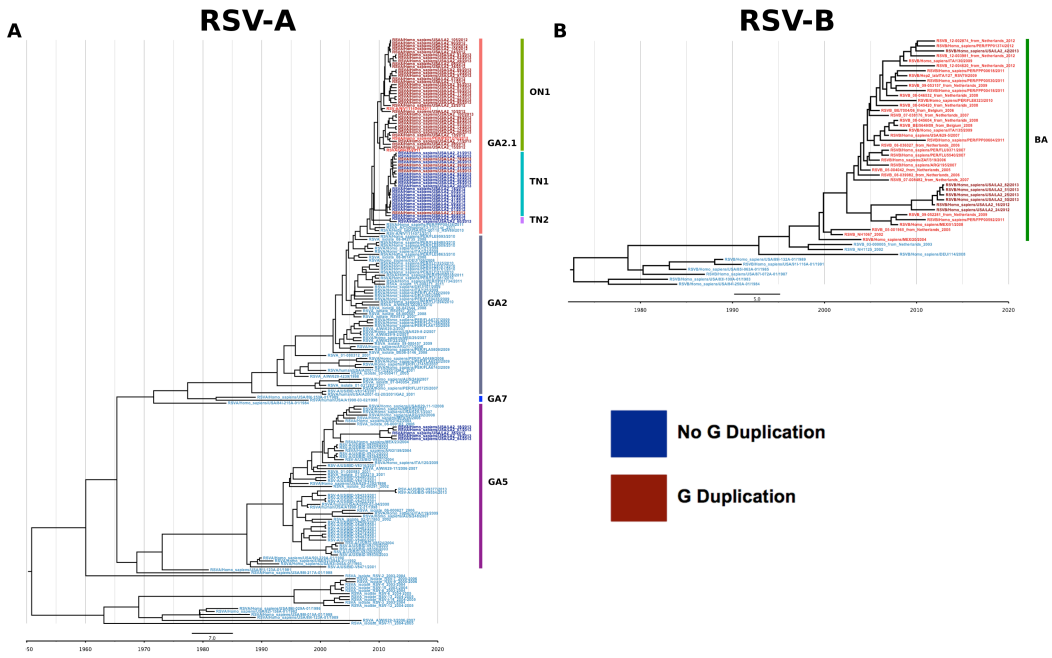


Figure 34 Bayesian Maximum Clade Credibility Trees for RSV-A (A) and RSV-B (B) G Gene Sequences. Strain names are colored by the presence (red) or absence (blue) of the large G gene duplication, with study samples in darker shades of red and blue. Multiple co-circulating lineages of RSV are observed during the 2012-2013 RSV season. These phylogenies and related analyses suggest that the G gene duplication occurred convergently in two separate genotypes of RSV-A.

mutation within the SH gene of RSV-B, similar to previous reports (43, 106). The Bayesian time to most recent common ancestor (tMRCA) mean estimates for the

whole-genome data set suggests circulating and historical RSV-A share a common ancestor from 1951, whereas available whole-genome RSV-B sequences likely diverged in 1967 (Table 10). Comparison of the RSV-B whole-genome phylogenies to the G gene phylogenies that contain more extensive sampling of all the available GenBank full G gene sequences (Figure 36) indicates that the whole-genome data set is missing the diversity that exists within several RSV-B G clades.

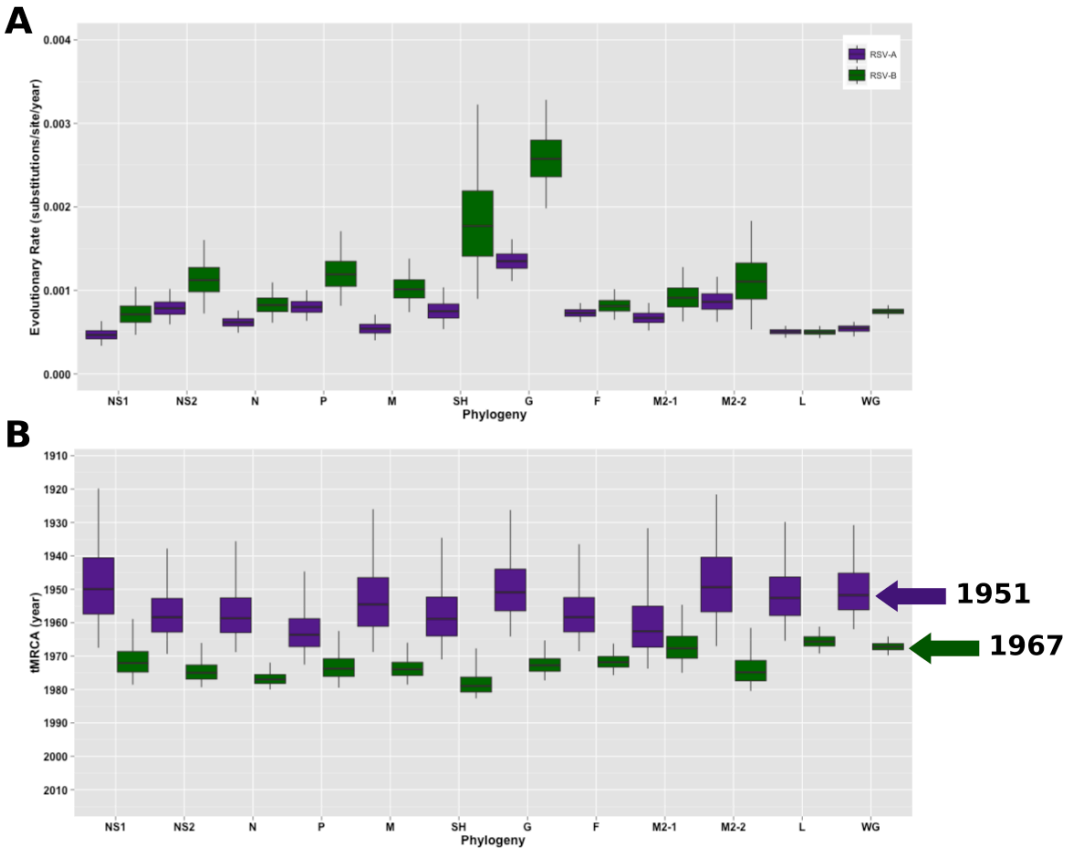


Figure 35 Times to Most Recent Common Ancestors (tMCRAs) and Mean Evolutionary Rate Estimates Inferred by Bayesian Analyses. Estimates are provided for RSV-A (purple) and RSV-B (green) for the whole genome (WG) and each individual gene. (**A**) Mean tMRCAs for RSV-A and RSV-B data sets and (**B**) evolutionary rates (substitutions/site/year) for RSV-A and RSV-B data sets are provided with 95% HPD intervals indicated over the box and whisker plots. The whiskers in each plot extend to the full 95% interval, the boxes indicate the 25-75% interquartile range of the posterior distribution, thus describing its central tendency. Mean whole-genome tMRCA estimates are indicated with arrows: 1951 for RSV-A and 1967 for RSV-B.

Table 10 Mean Evolutionary Rates (substitutions/site/year) and Times to Most Recent Common Ancestor (tMRCA), as Inferred by Bayesian Analysis

	tMRCA (95% HPD)	MeanRate (95% HPD)
RSV-A WG	1951 (1937-1964)	5.68×10^{-4} (6.55×10^{-4} to 4.87×10^{-4})
RSV-B WG	1967 (1964-1970)	7.47×10^{-4} (8.22×10^{-4} to 6.64×10^{-4})
RSV-A G	1949 (1928-1966)	1.35×10^{-3} (1.60×10^{-3} to 1.10×10^{-3})
RSV-B G	1972 (1966-1978)	2.59×10^{-3} (3.28×10^{-3} to 1.98×10^{-3})

WG = whole genome; G = G gene; HPD = highest posterior density

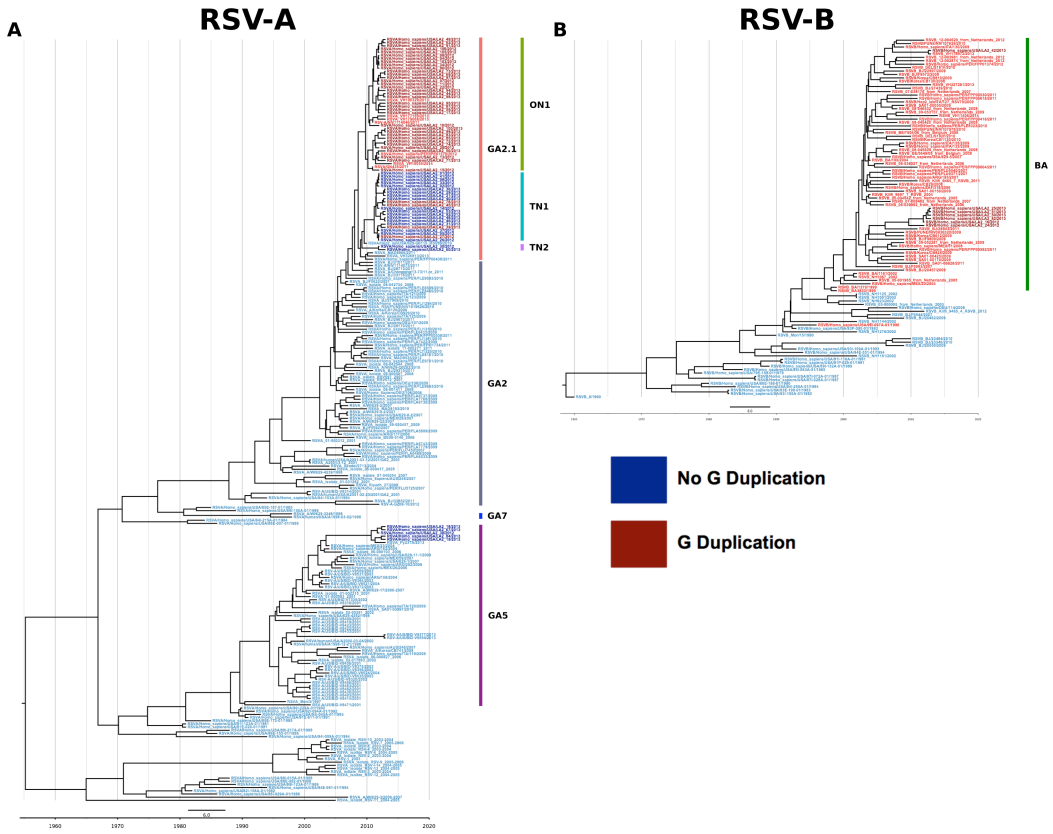


Figure 36 Bayesian Maximum Clade Credibility Trees for All Available Full G Gene Sequences Downloaded from GenBank and Down Sampled to Include Representative Centroid Sequences from 98% Sequence Identity Gene Clusters. (A) The RSV-A G gene phylogeny shows that relatively little new diversity is added compared with the whole-genome analyses. However, the RSV-B G gene phylogeny **(B)** shows that additional diversity is being sampled by including additional G gene sequences compared with the whole-genome analyses. This indicates better surveillance of RSV-B from G gene sequences than from whole-genome data set.

The Bayesian analysis of the G gene MCC phylogeny of RSV-A also supported the hypothesis that the G gene duplication occurred at least twice in a convergent manner within the RSV-A genotypes ON1 and TN1 (Figure 34A). This is

evident from the interleaving of the RSV-A G sequence duplicated genomes with the non-duplicated genomes within these genotypes in the G gene phylogeny, as well as by the divergence dating estimates for the recent GA2.1 strains and the genotypes contained within it: ON1, TN1, and TN2 (Figure 37). These results suggest that genotype ON1 diverged first in late 2009, followed by genotype TN1 in early 2011 (a local Tennessee clade), and finally by genotype TN2 in late 2011. Because the latter two genotypes appear to have evolved from a non-duplicated ancestral G gene sequence, genotype TN1 most likely acquired the duplication convergently. This hypothesis is also supported by the minimal overlap in the 95% HPD interval of the divergence time estimates for genotypes ON1 and TN1.

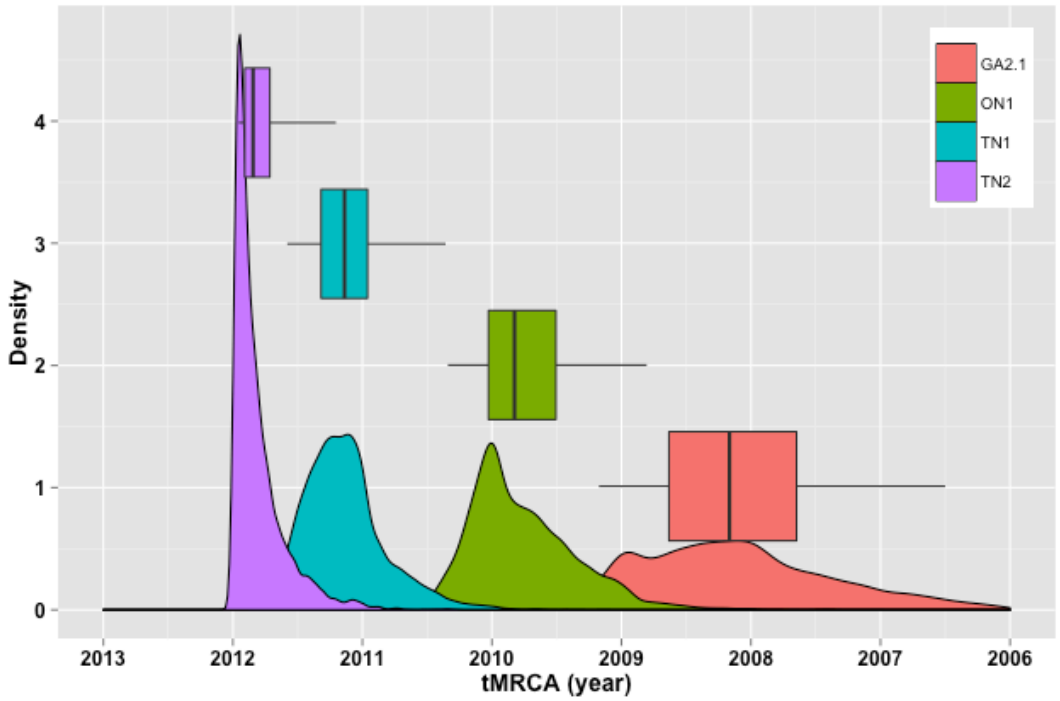


Figure 37 Divergence Time Estimates from a Bayesian Divergence Dating Analysis of the RSV-A G Gene Sequences. The GA2.1 clade consists of genotypes ON1 containing only sequences with the G gene duplication, TN1 containing sequences with mostly non-duplicated G genes and four interleaved G gene duplication sequences and TN2 containing only sequences lacking the G gene duplication. Divergence estimates suggest clade GA2.1 originated from a non-duplicated ancestor, with the duplication being convergently gained first in genotypes ON1 and then TN1. This hypothesis of convergent G gene duplications is supported by divergence estimates that largely do not overlap between genotypes ON1 and TN1.

Bayesian SkyGrid analyses indicate a change in population dynamics for both the RSV-A and the RSV-B viruses during the introduction and global spread of their respective G gene duplications (Figure 38). There is a population size reduction in 2004 for RSV-A and 2011 for RSV-B (Figure 38, purple and green arrows, respectively) followed by exponential growth, which can be seen in both the whole-genome and the G gene phylogenetic analyses.

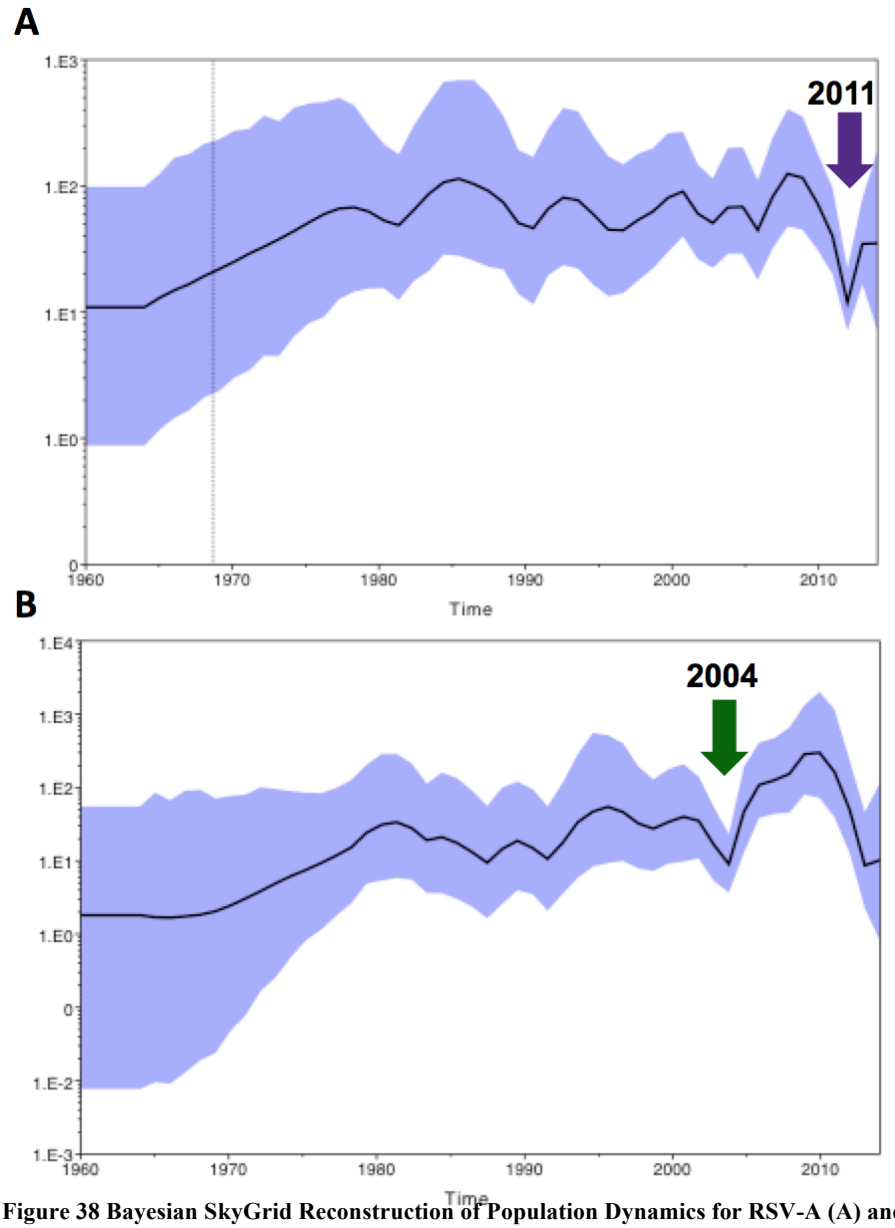


Figure 38 Bayesian SkyGrid Reconstruction of Population Dynamics for RSV-A (A) and RSV-B (B) G Gene Data Sets. The purple arrow indicates a bottleneck event in RSV-A corresponding to the G gene duplication entering global circulation and undergoing a subsequent population expansion. Similarly, the green arrow indicates the same phenomenon occurring around the time that the RSV-B G gene duplication reached global predominance. These data suggest that the G gene duplication provided RSV-A and RSV-B with selective advantages that allowed them to spread globally and replace previously circulating lineages.

Glycosylation Analysis Reveals Genotype Specific Glycosylation Patterns in the G Protein

Results of NetNGlyc across all study samples showed that the N-linked glycosylation in the F gene was relatively conserved. Nearly all of the RSV-A and RSV-B samples had the same N-linked sites (27, 70, 116, 120 and 126) within the F2 domain. N-linked glycosylation on the G gene appeared to follow a genotype specific pattern, with multiple glycosylation patterns co-circulating simultaneously. Genotype ON1 had three predicted N-linked glycosylation sites, clade TN1 had five sites, clade TN2 had four sites, and our study viruses from GA5 had five sites (Figure 39). RSV-B genomes showed two different glycosylation patterns. Genotype BA.1 had four glycosylation sites, three of which were consistent with the majority of circulating

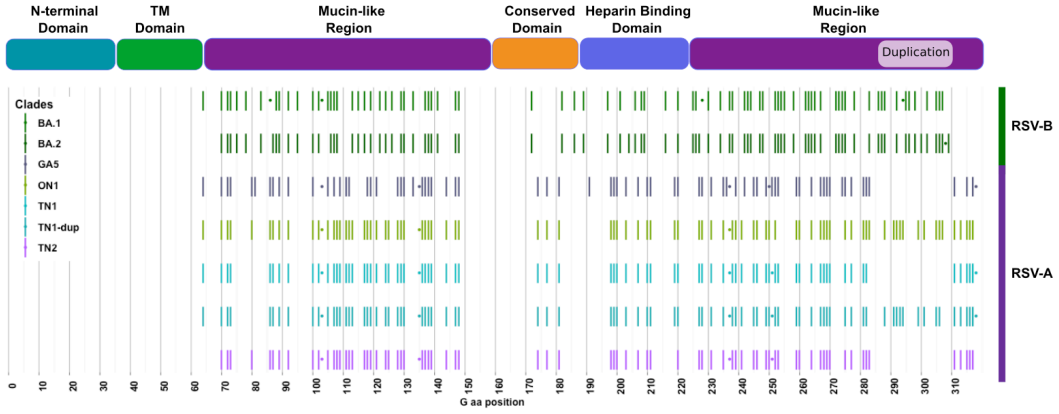


Figure 39 Consensus N- and O-Linked Glycosylation Patterns for the Seven Study Genotypes. The seven genotype specific consensus glycosylation patterns for O and N-linked (bars and dots respectively) glycans are displayed in rows. RSV-A and RSV-B genotypes are indicated with purple and green bars to the right. Each genotype displays a unique glycosylation pattern and duplication status

RSV-B genomes. Genotype BA.2 showed a novel RSV-B glycosylation pattern with just one glycosylation site present toward the C-terminal end of the G protein after the G duplication. Similarly, NetOGlyc showed that the O-linked glycosylation patterns for the G protein followed a genotype specific pattern as well. Seven distinct O-linked patterns were observed in the G protein sequences from our study cohort. Genotype ON1 showed 85 predicted O-linked glycosylation sites, whereas genotype TN1 had 74 and 83 sites (in non-duplicated and duplicated genomes, respectively), genotype TN2 had 74 sites, and genotype GA5 had 72 sites. The RSV-B genomes showed two different glycosylation patterns: 82 sites for genotype BA.1 and 85 sites for genotype BA.2. There were no significant numbers of O-linked glycans predicted for the F protein. Consensus genotype specific glycosylation patterns were plotted for visual analysis for the G protein (Figure 39).

Table 11 Observed Indels and Start and Stop site Variants within RSV-A, RSV-B and Between A and B Groups. More indels are observed within the RSV-B group, particularly in the G gene suggesting greater plasticity of G in RSV-B. Additionally, RSV-A -B differences, especially in L suggest the potential for functional difference in the polymerase that may lead to this apparent greater rate of indels in -B.

Group	Gene	Indel
RSV-A	F	none
	G	**780-852**, 966-stop, 969-stop
	L	400-406
	M	none
	M2-1	none
	M2-2	1-start, 7-start
	N	none
	NS1	none
	NS2	none
	P	none
	SH	none
RSV-B	F	none
	G	471-477, 673-685, 704-707, **793-853**, 954-stop, 963-stop, 975-stop
	L	none
	M	none
	M2-1	none
	M2-2	1-start, 10-start
	N	none
	NS1	none
	NS2	none
	P	none
	SH	none
Inter Group	F	none
	G	471-477, **634-637**, 673-685, 704-707, **793-853**, **857-929**, **1003-1008**, 1029-stop, 1038-stop, 1042-stop, 1045-stop, 1050-stop
	L	400-406, **5193-5196**, **5280-5283**, **6504-stop**, **6507-stop**
	M	none
	M2-1	**582-stop**, **585-stop**
	M2-2	1-start, 10-start, 16-start
	N	none
	NS1	none
	NS2	none
	P	none
	SH	**180-183**

Gene Sequence Plasticity Contributes to Variability Between and Within RSV Groups

Analysis of gene alignments within and between RSV groups showed various indels (especially in the G gene) as well as various start site and stop site variant sequences (Table 11). In RSV-A, we observed one indel each in the G and L genes, and two start site variants in the M2-2 gene. There were also two stop site variants in the G gene data set. Within the RSV-B data set we observed four indels in the G gene with three stop site variants. We also observed two start site variants in the M2-2 gene. Comparing RSV-A to RSVB we observed an additional seven indels: two in the G gene, four in the L gene, and one in the SH gene. In addition to these intergroup indels, there were two variant stop sites found in

the M2-1 gene between groups. Interestingly, one M2-2 start site variant was shared between subtypes, whereas one each was unique in RSV-A and RSV-B, leading to only three sites observed in the intergroup comparison, although none were novel observations.

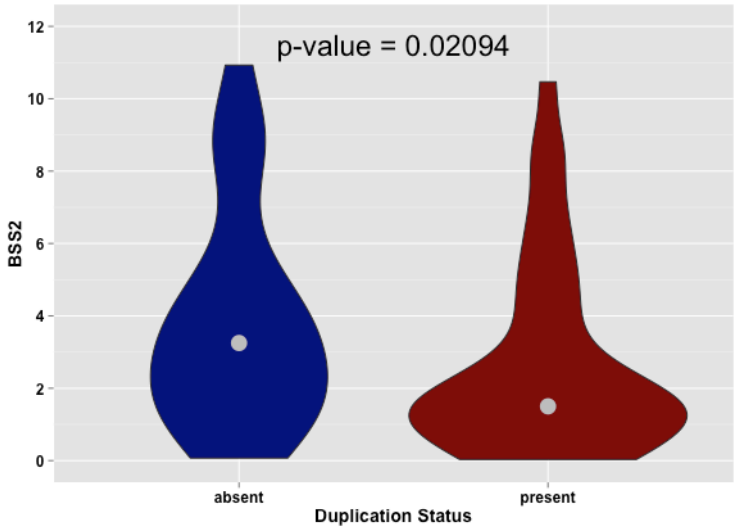


Figure 40 Comparison of Bronchiolitis Severity Scores (BSS2) with the Presence or Absence of the 72-nucleotide RSV-A G Gene Duplication. A violin plot of severity scores by duplication status shows lower median BSS2 (indicated with grey dots) in patients with the G gene duplication. The shape of the violin plot shows this effect is mainly observed for lower BSS range, suggesting that other factors (e.g., host immune response or microbiome) also contribute significant roles in overall disease severity.

Detection of Global and Local Circulation Patterns Ssing BaTS Analysis

Bayesian Tip-association Significance (BaTS) testing of the RSV-A G gene phylogeny resulted in AI and PS scores of 0.0, and MC scores of 0.009 for the global and local state assignments. These results are indicative of strong evidence for these states being topologically associated with the Bayesian phylogenies.

RSV-A 72-Nucleotide G Gene Duplication was Associated with Reduced Disease Severity in Infants

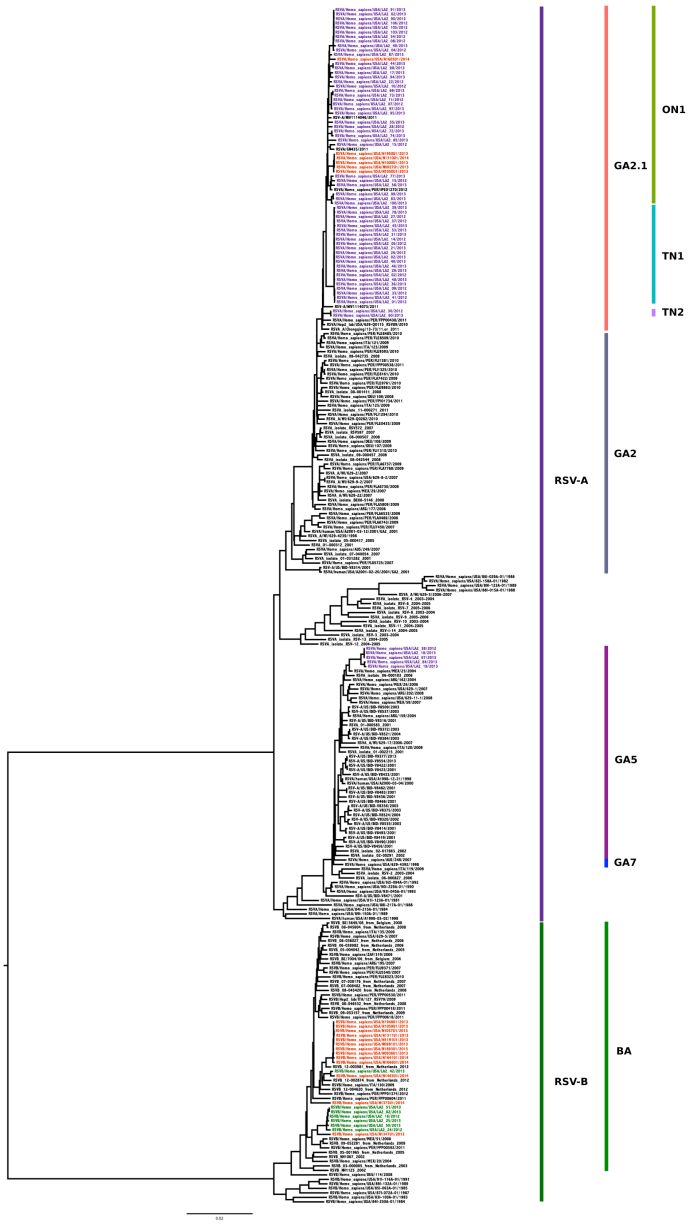


Figure 41 Maximum Likelihood Phylogeny of a Pruned RSV Whole-Genome Data Set. The data set for this phylogeny is the same as for Figure 29, with the addition of 20 genomes collected during the INSPIRE study in Tennessee during the 2013-2014 RSV season. RSV-A study isolates are depicted in purple, whereas RSV-B study isolates are depicted in green; the 2013-2014 genomes are depicted in orange. The 2013-2014 season in Tennessee witnessed a switch to RSV-B dominance, as well as the continued circulation of RSV-A genomes containing the G gene duplication (6 out of 6 RSV-A 2013-2014 study strains).

We assessed the relationship between the C-terminal of the G sequence duplications and the clinical severity data via the bronchiolitis severity score 2 (BSS2) metric, using Wilcox-rank sum test. The BSS2 score measures the severity of lung involvement with RSV infections by assessing infants on the following criteria: respiratory rate, pulse rate, retractions, dyspnea, and auscultation (125).

Within infants whose samples had the RSV-A strain, those without C-terminal G sequence duplication had significantly higher BSS2 values (p=0.2094) when

compared with infants with the duplication (Figure 40). These results suggest that there is a difference in mean severity between the disease caused by the viruses with the duplication and those without – severity is higher in the older viruses without the duplication. Also, a difference was seen between the proportion of male and female hosts infected with viruses with and without the G gene duplication, however this result was not significant.

We performed a whole-genome phylogenetic analysis using the 2012-2013 data set along with a limited number of 2013-2014 genome sequences. The resulting maximum likelihood phylogeny (Figure 41) showed a switch from RSV-A to RSV-B predominance. We also noted the continued exclusive circulation of RSV-A ON1 genotype viruses, all of which contain the G gene duplication.

Discussion

Here, we have identified multiple co-circulating RSV variants infecting infants within the central Tennessee region during the 2012-2013 season. Substantial RSV genetic diversity was observed during the 2012-2013 RSV season in central Tennessee, both between and within the RSV-A and -B groups. This diversity was especially evident within the G gene, although additional evidence was present in the F and select regions of the L gene. We observed seven distinct G gene genetic variants in our data set, as defined by both G gene duplication and glycosylation status. This observed diversity is likely part of the mechanism RSV uses to evolve and evade host adaptive immune responses (126-128). We have also described a possible phenotype for the recent G gene duplications seen in RSV-A and RSV-B. A previous study made and association between the RSV-B G gene duplication and

enhanced viral attachment (129). Our data suggests an association between the G gene duplication in RSV-A and lowered disease severity, as measured through BSS2.

These data demonstrate that during the 2012-2013 RSV season, three distinct lineages of RSV (each containing multiple genomes) were co-circulating within the central Tennessee region. This supports previous reports of multiple RSV types co-circulating (109, 111, 130). Furthermore, RSV-A clade GA2.1 appears to have predominated, accounting for 83.1% of the observed infections in our study cohort. In this study, we observed a local subgroup of RSV-A within clade GA2.1 that appears to be circulating only in central Tennessee (genotype TN1). The continued circulation of the ON1 genotype viruses during the 2013-2014 season supports the hypothesis that the 72-nucleotide duplication within the G gene is moving toward fixation in the RSV-A population. Missing sequence diversity in the RSV-B whole-genome data set may explain the relatively large ranges obtained for the Bayesian substitution rate estimates for many of the genes compared with those for RSV-A. The population dynamic analyses using Bayesian SkyGrid plots indicate that the introduction of the G gene duplication into RSV genomes reduced viral genetic variability (shown as a bottleneck event), which was quickly followed by an exponential expansion of the population size, suggesting a fitness advantage for the duplication variants.

Comparing publicly available whole-genome RSV sequences to all available full-length G gene sequences indicated that, although whole RSV-A genomes are largely representative of known RSV-A diversity, the corresponding RSV-B whole-genome data set is missing diversity within the RSV-B BA clade. This finding

suggests that additional whole-genome sequencing of historical samples would greatly improve our understanding of RSV-B diversity. Similarly, a comparison of RSV-A and RSV-B whole-genome sequences shows that RSV-B contains more insertions and deletions (indels) within the G gene, suggesting that different selection pressures exist between these groups. Overall higher mutation rates in RSV-B and specifically the difference in the SH gene mutation rates between RSV-B and RSV-A further support this, although this could also be a result of poor RSV-B surveillance. The fact that intergroup indels occur in genomic regions other than the G gene supports the need to adopt a whole-genome sequencing approach for future RSV studies.

In this study, we observed tree topologies with little or no temporal and or geographic patterns and others with strong geographic and temporal patterns. Over the long-term, most genetic diversity within both RSV-A and RSV-B appears to circulate globally on relatively short time scales. However, in our study and other studies, such as Agoti et al. (43), localized RSV evolution is sometimes apparent. Genotype TN1 appears to be a central Tennessee RSV-A cluster, whereas several groups of RSV-B that were noted to be local to Kilifi, Kenya in the Agoti study (43). We tested the assumption of localized clades using BaTS with a significant result, suggesting that genotype TN1 and Tennessee GA5 viruses were being locally transmitted during the 2012-2013 season. With broader genomic surveillance of RSV, these epidemiological patterns can be studied more closely and the origins of various lineages could be determined.

With the addition of 7 new RSV-B whole-genomes (11% of total publicly available RSV-B whole-genome sequences), our tMRCA estimate for RSV-B is likely improved over previous estimates (106); however, the estimate would likely be improved with additional whole-genome sequences of historical RSV-B genomes since RSV-B was first identified in 1960.

The convergent appearances of large G gene duplications in the same genome location for multiple RSV lineages suggest that the G protein is highly permissive for these types of insertions and that a specific replication mechanism is responsible for their generation. These duplications demonstrate the plasticity of the G protein and its tolerance for insertions, and they potentially indicate a mechanism for the development of novel immune evasion strategies. We observed just one large 72nt duplication in the G protein of RSV-A, as well as two stop-variants, and four indels of various sizes with three stop-variants in the RSV-B G protein data set. The existence of two major duplications may also indicate that the duplications imparts some level of selective advantage for the virus, because the duplication appears to have reached fixation in RSV-B genomes and may be moving toward fixation in RSV-A, although improved RSV surveillance/sampling is required to know this for sure. A recent study by *Hotard et al.* showed an association between the 60nt G gene duplication in RSV-B and an enhancement of the attachment function of the G protein (129). It is possible that the 72nt duplication in RSV-A similarly enhances the G protein, thus providing a selective advantage to duplicated viruses. As previously reported, there may be a mutation in the G gene that primes duplication to occur (113-115), making it more likely to happen repeatedly. It has been proposed

that stem loop structures form in the replicating RNA strand, causing the polymerase to pause and reinitiate replication further back on the template (110). The apparent observation of this duplication occurring repeatedly in RSV-A, and the duplication being of the same length and location, supports this proposed mechanism as an explanation for the duplication events. From an evolutionary perspective, coalescent theory suggest that most currently observed genomes arose for a relative few ancestral sequences. Similarly, most contemporary genomes will not continue on to become ancestors of future viruses. Any given transmission chain is likely to die out; thus, repeated introductions of the same mutation makes it more likely that advantageous mutationss such as the G gene duplications will begin to circulate globally. The observation that the local Tennessee genotype, TN1, didn't reach global circulation supports this notion, although poor global surveillance of RSV is an alternative explanation. Similarly, at least one RSV-B duplicated genome exists in our data set from 1996, earlier than the previous first observation in 1999 of a 60nt duplication in the G gene of RSV-B (114). It should be noted that the 1996 genome was located in a separate clade from the BA clade where RSV-B duplicated genomes originate, supporting convergence as a mechanism for the increased probability of the success of these mutations in reaching global circulation.

Our results suggest that, although the G gene duplication does appear to influence disease severity, it is likely a relatively small effect. RSV severity, as has been reported elsewhere (85, 104-106), is largely influenced by known risk factors, such as prematurity, gender, chronic lung and heart diseases, and immune deficiencies. Furthermore, it is likely that additional, less well-characterized host

genetic factors, nasopharyngeal and upper respiratory microbiomes and viral genetic characteristics all play significant roles in the development of severe RSV disease. Future studies that incorporate either genome-wide host genetic associations or targeted genetic analysis (e.g., of the HLA region) with microbiome and viral genetic information will be important in developing models that elucidate the interplay between the host, the microbiome and the pathogen in severe RSV disease progression.

In general, glycosylation patterns in the RSV-B data set appear more varied than in the RSV-A dataset (data not shown), which is supported by the idea that RSV-B seems to have more G gene plasticity than RSV-A. These results reveal a relatively static F protein, in terms of glycosylation, compared to a dynamically glycosylated G protein that swaps in and out different glycosylation sites relatively quickly. These differences in G glycosylation may help the virus spread and overcome herd immunity, in addition to the other differences noted in the G protein.

The high degree of conservation in the F1 domain likely is required to maintain a functional fusion mechanism, as this protein undergoes a complex conformational change once attachment triggers fusion (131). This function would likely be greatly hindered by variability within F1, possibly due to steric hindrance (131). The F protein has a conserved glycosylation pattern across RSV-A and RSV-B viruses and appears to only permit N-linked glycans in the F2 domain. Again, this may be due to the need for F1 to be sterically free to change confirmation to perform its fusion function. The overall conservation of the F gene juxtaposed against the

variability of the G gene suggests that the F gene is a more suitable target for universal therapeutics and vaccines (132).

One major limitation of this study was the lack of extensive historical and contemporary sampling to place our whole-genome sequences in context. For instance, without a robust surveillance network for RSV, it is hard to know for sure if the TN1 genotype was truly geographically isolated to Tennessee during the 2012-2013 season. Another notable limitation is the relatively small sample sizes with which to perform the statistical associations of clinical and genetic data. Both point toward the need for expanded surveillance and coordination of clinical data collection.

Conclusions

Here we have identified several co-circulating RSV variants infecting infants within the central Tennessee region during the 2012-2013 infectivity season. Differing selective pressures and variation patterns were apparent in our data set suggesting epidemiological and evolutionary mechanisms for immune evasion. For the first time, possible phenotypes for the recent G gene duplications seen in RSV-A have been described. We have associated the G gene duplication in RSV-A with a trend toward decreased pathogenicity. Further investigation of these patterns and surveillance of RSV is called for. Additionally, studies that aid the identification of further therapeutic and vaccine targets would be beneficial.

Chapter 5: The Future of Viral Comparative Genomics

The work described above is an example of a new approach to data-driven bioinformatic research that aims to integrate experimental results, from sequencing to phenotype analyses, with metadata describing those data in a new and meaningful way. In study one, we have described a novel method that extends the current capabilities of antigenic distancing techniques by directly connecting sequence-based hierarchical clusters with phenotypic measurements. We have incorporated bootstrap analysis into these phenotype/genotype correlations to take into account the variability in the antigenic measurement, and hence to provide a more robust estimation of the reliability of our predictions than has previously been possible. By cross-referencing our results with the established phylogenetically defined clades, this technique establishes an even more powerful method for conducting viral sequence and antigenic surveillance – in addition its valuable vaccine selection function. Consequently, our methods point toward a future of antigenic surveillance where all the viral samples collected are first sequenced and then this sequence data is used to determine which viruses to perform antigenic characterization on. One notable extension to the analysis described here would be to perform similar predictions on especially important avian influenza subtypes with human pandemic potential, such as H5N1 and H7N9. Additionally, the methods described in this study could conceivably be applied to other viral species for which antigenic characterization is needed. Similarly, assays that generate data pertaining to other phenotypes could potentially be used to predict the extent in sequence space that those phenotypes extend.

In our second study, we extended the current offerings of methods for performing viral genome constellation analysis. We described VirComp and OrionPlot as two separate, but related, pieces of software for performing constellation analysis based on gene clusters and providing publication ready figures. The creation of OrionPlot greatly extends the abilities of researchers to publish this type of genotype analysis, as no similar tools currently exist for generating constellation figures so quickly. We described one application of VirComp and OrionPlot outside gene-based constellation analysis: it was used to uncover methodological errors in the amplicon-based sequencing approach that lead to the discovery of a recombinant virus false positive. This establishes these tools as excellent candidates for incorporation into sequencing quality control analysis pipelines for viral or other amplicon-based sequencing projects. In addition to this use, VirComp and OrionPlot could easily be incorporated into the viral the BRCs, ViPR and IRD, as a novel comparative genomics tool. The methodology could be used to perform analysis on non-viral genes as well. It would be most useful on small groups of related genes, possibly in metabolic pathways, or perhaps multi-locus sequence type (MLST) genes. OrionPlot could conceivable be used as a stand-alone analysis program to generate visualizations of categorical metadata. Although the current implementation of OrionPlot requires a data matrix with numerical categories, an adjustment to the code to allow for character data instead of numerical data would allow OrionPlot figures to be made for various categorical metadata. Plots of metadata could then be combined with phylogenies to provide powerful visuals showing the associations of the metadata to the sequence analyses.

In study three, we directly examined the relationship between genetic variants and phylogenies with clinical metadata gathered from infants infected with RSV. We showed that multiple lineages of RSV-A and -B co-circulated in central Tennessee during the 2012-2013 RSV season. These co-circulating lineages are defined by distinct O and N-linked glycosylation patterns on the G protein. Some of the lineages appeared to be locally transmitting within our study region, whereas others were globally circulating. One lineage in particular was shown to contain a duplication within the G protein. This duplication appears to have arisen in a convergent fashion within its global lineage and with four members of a locally circulating lineage. Through statistical associations with the clinical metadata collected for this study, we have shown an increased number of infections of infant females with viruses having genomes containing the G duplication, as compared to those without the G duplication. Similarly, we have shown a decrease in the severity of the disease, as measured via BSS2, in hosts infected with the G gene duplicated genomes. Although the overall effect of this virus-derived genetic signature of severity was small, it suggests avenues of research to examine specific mutations that may be associated with particular viral phenotypes. Additionally, expanding the scope of this study to integrate heterogeneous data types beyond viral genetic data and clinical metadata could provide further insights into the interplay between host, pathogen and environment that determine the course and outcome of an infection.

Viral whole-genome studies will undoubtedly play a large role in the future of viral research, from basic research into viral pathology and epidemiology, to novel global viral surveillance, and the development of vaccines and other viral therapeutic

drugs. The proliferation of large-scale sequencing efforts in viral cohort studies is of the utmost importance in transforming viral research into a science capable of benefiting from the enhancements of big data. To that end the near term requires the development of both tools and standards in practice for conducting viral research in an age of big data. Efforts to standardize metadata, such as those by the GSCID-BRC Metadata Working Group (31), are key first steps in achieving data integration. An example of the type of data standard that is necessary is controlled vocabularies (or ontologies), such as those available from the Open Biomedical Ontologies Foundry (OBO) (133). It should be noted that resources such as the OBO Foundry require ongoing curation as new data types come into use and older ones are retired or refined. Similarly, use of these ontologies requires data expertise. Tools for expanding viral genomics into an age of big data could include technologies such as representational state transfer (RESTful) web services. Through RESTful interfaces labs across a spectrum of biological or clinical settings could expose and integrate their experimental data types (134) in near-real time, and to gain access to computational resources (135). This would have applications for expediting the processing of viral outbreak surveillance data, emergency management, and the development of vaccines and therapeutics.

The integration of viral sequence data with, host sequence data (either whole genome or targeted to specific host-pathogen interfaces), and environmental sequence data (such as microbiome or metagenomic data) is on the cusp of becoming commonplace. Integrating these disparate data sources in meaningful ways will require the novel application of advanced statistical approaches and the development

of visualization and algorithmic tools (136, 137). For instance, sequencing the host HLA region of RSV-infected subjects has the potential to establish immune related host factors that may determine the severity of infection (132). Similarly, the metagenomic characterization of RSV-infected individuals would provide an additional dimension to the understanding the pathogenesis of the virus. Understanding the interplay between host, environment (microbiome or metadata), and pathogen will require novel applications of statistical models to adequately assess these relationships. We have described a data exploration technique above known as MDS, which falls into a class of exploratory techniques known as ordination (138). Ordination techniques also include principal component analysis (PCA) (138). We have also described hierarchical clustering techniques for performing sequence analysis, such as Wards's or farthest neighbor clustering (74, 102). Other methods of clustering, such as k-means, will also continue to be a useful data exploration technique for heterogeneous data (138). In the recent NCI-DREAM experiment, candidate chemotherapies were assessed using a variety of statistical prediction models (139). These models included: support vector machines, linear and non-linear regression models, and ensemble models (the use of multiple models composited together for a more informative result) (139). Many of these techniques fall into the category of machine learning algorithms that are used to make predictive associations. Needless to say, there are countless approaches to modeling that could be applied to viral comparative genomics using big data. These types of approaches will hopefully produce novel insights into the pathogenic and epidemiologic processes in viral research using mixed data.

Public health policy has and will continue to be influenced by heterogeneous data. For instance, as described above, the WHO utilizes both antigenic characterizations and sequence data to determine the selection of vaccine candidates for the annual influenza vaccine. Our methods will certainly enhance the integration of these data to establish a more informative vaccine selection process. Epidemiological modeling has been suggested as a target for further integration of data types to aid in developing public policy around specific infectious disease agents (140). An example of this is the integration of social media reports of disease with public data relating to the epidemiological progress to create real-time risk maps of infections based on geographical location (137). These integrated data will certainly aid in development of the public policies surrounding viral pathogens.

The work described in this dissertation and the directions laid out above are but a few of the myriad of efforts to integrate data, compare viral pathogens, and make inferences about pathogenesis and epidemiology. These efforts aim to glean new insights into the world of infectious disease and to better understand the interplay between the host, the pathogen, and the environment. Through this research, novel vaccine targets and therapeutics will undoubtedly be identified. Big data approaches to biological research are certainly the future of the science and the need for data integration and visualization tools is clear.

Appendices

Appendix A. Interim PTP and DASH report of Influenza activity from December 2013

December 1, 2013 Interim Manufacturing-at-Risk (MAR) Candidate Selection Report

JCVI NFLU Team

January 22, 2014

Summary of MAR Candidates

The following is a summary of MAR candidate viruses based on DASH and PTP analysis. A detailed report of the analyses that produced these results follows Table 12 and is broken down by Influenza A subtype and Influenza B lineage

Table 12 Summary of MAR Candidates with Priority and Rationale.

Strain Name	Type	Priority	Rationale
A/Almaty/2958/2013	H3N2	High	WHO candidate virus
A/Estonia/76676/2013	H3N2	High	High scoring DASH vaccine candidate
A/Norway/2255/2013	H3N2	High	High scoring DASH vaccine candidate
A/Estonia/76614/2013	H3N2	Medium	High scoring DASH vaccine candidate
A/Cameroon/12V-5136/2012	H3N2	Medium	High scoring DASH vaccine candidate
A/England/358/2013	H1N1pdm	High	High scoring DASH vaccine candidate
A/Bolivia/559/2013	H1N1pdm	High	WHO candidate virus
A/Estonia/74816/2013	H1N1pdm	Medium	High scoring DASH vaccine candidate
A/Dominican Republic/7293/2013	H1N1pdm	Medium	WHO candidate virus
B/Massachusetts/2/2012*	Influenza B Yam	High	High scoring DASH vaccine candidate
B/Lithuania/6942/2013	Influenza B Yam	Low	High scoring DASH vaccine candidate

B/Belgium/G886/2012	Influenza B Vic	High	High scoring DASH vaccine candidate
B/Texas/2/2013	Influenza B Vic	High	High scoring DASH vaccine candidate
B/Formosa Province/V2367/2012	Influenza B Vic	High	High scoring DASH vaccine candidate
A/Colorado/21/2012	H3N2	Low	DASH gap filling MAR candidate
A/Wisconsin/6/2013	H3N2	Low	DASH gap filling MAR candidate
A/New York/3214/2013	H3N2	Low	DASH gap filling MAR candidate
A/Helsinki/823/2013	H3N2	Low	DASH gap filling MAR candidate
A/Puerto Rico/1/2013	H1N1pdm	Low	DASH gap filling MAR candidate
A/Florida/43/2013	H1N1pdm	Low	DASH gap filling MAR candidate
A/New Hampshire/4/2013	H1N1pdm	Low	DASH gap filling MAR candidate
A/Minnesota/26/2012	H1N1pdm	Low	DASH gap filling MAR candidate
A/Nizhny Novgorod/RII02/2013	H1N1pdm	Low	DASH gap filling MAR candidate
B/Iowa/4/2013	Influenza B Vic	Low	DASH gap filling MAR candidate
B/Sao Paulo/2-22035/2013	Influenza B Vic	Low	DASH gap filling MAR candidate
B/Indonesia/Nihrd-Buas704/2013	Influenza B Vic	Low	DASH gap filling MAR candidate

*Current Vaccine Seed

Proportion Tracking Pipeline (PTP) Analysis

The proportion tracking graphs depict bars that represent partitions windowed across three seasons. Each partition is weighted for the three seasons, e.g., the windowed partition "2012_11" represents data from three seasons viz. from November 2012 through April 2013, from May 2012 through October 2012, and from November 2011 through April 2012, with weighting factors of 0.50, 0.33, and 0.16, respectively. Thus, we achieve a sliding window (spanning 3 seasons) that slides across the seasons at the rate of one season per partition. Each partition is split into its component clusters based on sequence similarity and the arrows across partitions indicate similarity between the clusters they connect. A strong (bolder) arrow suggests a stronger sequence similarity between the clusters.

The clusters with colored circles contain and track the vaccine seed recommendations across partitions. A green circle indicates that the vaccine seed recommendation remained unchanged from the previous partition; a red circle represents a change in vaccine seed recommendation during the most recent time slice represented within that partition.

Distancing of Antigenicity by Sequence-based Hierarchical Clustering (DASH) Analysis

DASH analysis was performed using sequence data from strains collected during the period between December 2, 2012 and December 1, 2013. The HI data used was collected between December 2, 2012 and December 1, 2013. This selection of HI data includes February 2013 tables from NIMR, as well as February 2013 tables from CDC presented at the VRBPAC meeting on February 27, 2013. September 2013 NIMR HI tables have also been processed and included in the analysis for this reporting period.

H3N2 Analyses

H3N2 Proportion Tracking

Windowed proportion tracking analysis was performed on H3N2 sequences isolated between January 1, 2001 and December 1, 2013 using the HA1 domain. The current vaccine recommendation, A/Texas/50/2012), falls in the smaller diminishing portion (cluster 3) for the current partition (2013_11). The cluster that the current vaccine was in the last report, accounting for 93% of the sequences, has split into three different clusters with 45% in cluster 1, 40% in cluster 2 and only 13% in cluster 3,

which has the current vaccine recommendation. Cluster 4, which accounts for only 2% of the sequences, is composed of H3N2v viruses (Swine Origin Influenza Viruses – SOIV).

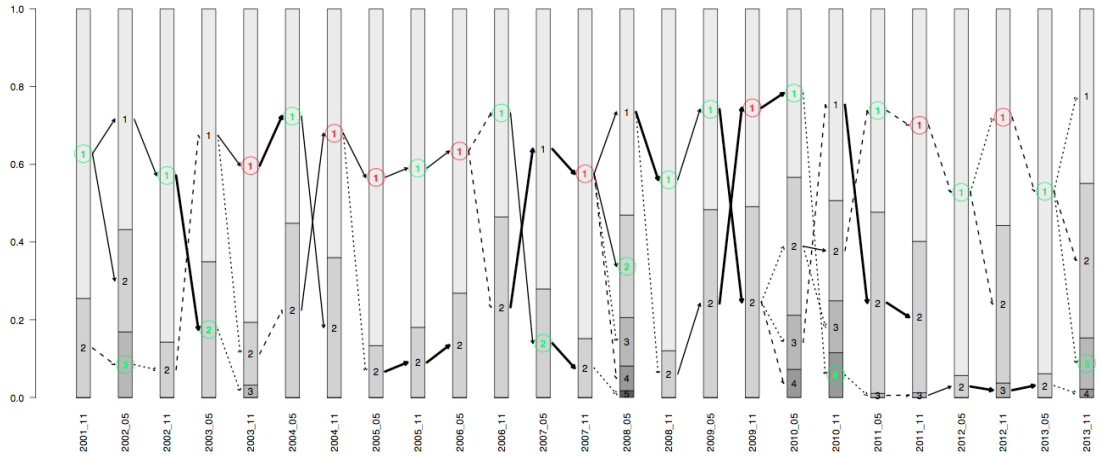


Figure 42 Proportional Sizes of H3N2 Clusters as Fraction of Total Sequences.

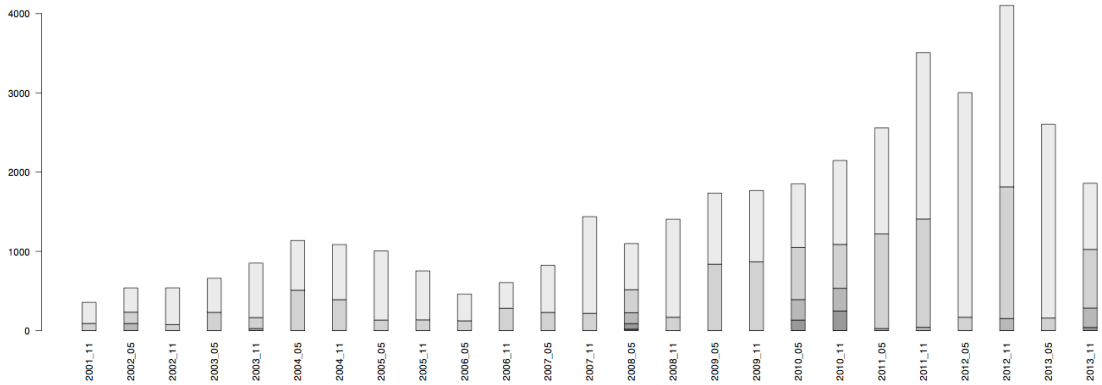


Figure 43 Sequence Counts Used for H3N2 Windowed Proportion Tracking.

H3N2 Proportion Tracking without Swine Origin Viruses

Windowed proportion tracking analysis was performed on all non-swine origin H3N2 sequences isolated between Jan 1, 2001 and December 1, 2013 using the HA1 domain. As shown in Figure 3, the current vaccine recommendation,

A/Texas/50/2012, falls in the smallest cluster (3) for the current partition (2013_11) accounting for only 10.6% of the circulating sequences of the total H3N2 sequences.

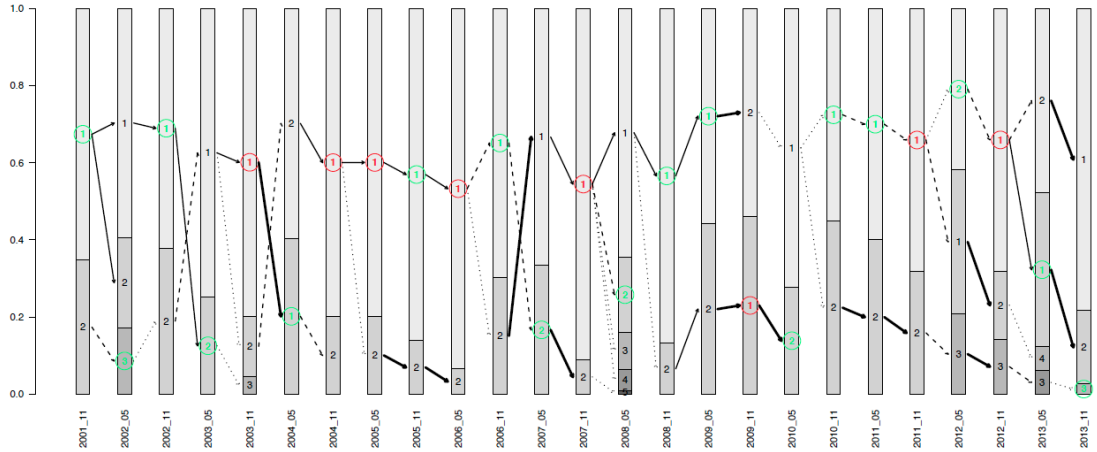


Figure 44 Proportional Sizes of H3N2 Clusters as Fraction of Total Sequences (Without Swine Origin Viruses).

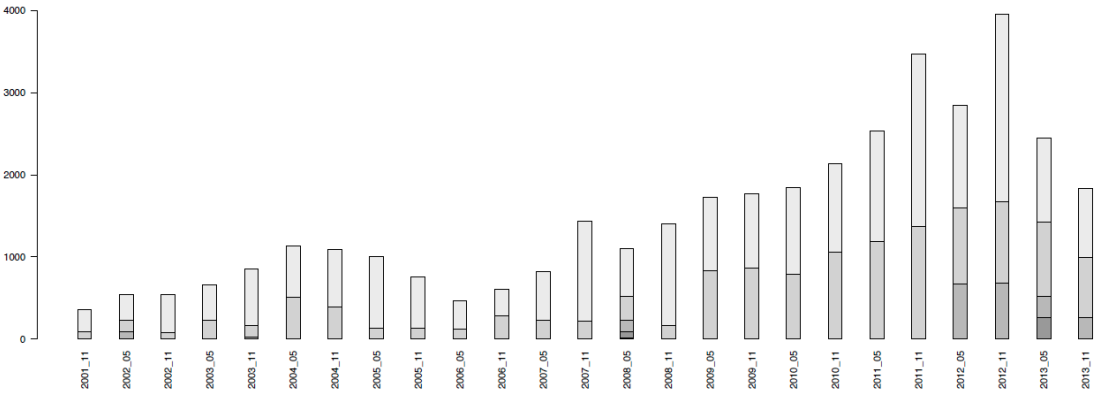


Figure 45 Sequence Counts Used for H3N2 Windowed Proportion Tracking (Without Swine Origin Viruses).

DASH – MAR Candidate Selection

Analysis of DASH results for both clusters without HI cohorts and candidate antigenic coverage prediction produced the following list of antigens that should be constructed as vaccine candidates or viruses to fill holes in HI antigen data.

PTP cluster 1 candidates (Clade 3C-3):

A/Almaty/2958/2013 (WHO candidate virus)

A/Estonia/76676/2013

A/Norway/2255/2013

PTP cluster 2 candidates (Clade 3C-2):

A/Estonia/76614/2013

A/Cameroon/12V-5136/2012

DASH gap filling viruses:

A/Colorado/21/2012

A/Wisconsin/6/2013

A/New York/3214/2013

A/Helsinki/823/2013

At an antigenic distance of 2, A/Victoria/361/2012 cell-propagated virus has a predicted coverage of circulating viruses of 46.0%. In our analysis, the A/Texas/50/2012 cell-propagated virus is predicted to cover 10.8% of circulating viruses. The Victoria/361 and Texas/50 egg-propagated viruses are predicted to cover 45.8% and 16.7% of circulating viruses respectively. This is a rather poor result for both Texas/50 viruses and suggests the WHO recommendation of the Texas/50 vaccine seed may not produce a noticeable difference in vaccine

effectiveness. It should be noted that the predicted antigenic coverage for all four viruses remains rather poor. Furthermore, the best vaccine candidates available, listed above, are only predicted to cover currently circulating viruses at around 70-85%. The apparent reason for the poor showing by the current vaccine candidates is that they poorly cover viruses from two of the subclades with the highest expansion rate; 3C-2 and 3C-3, while maintaining adequate coverage of their own clade 3C-1 as well as clade 5, 6, and 3A viruses. Conversely, A/Estonia/76676/2013 and A/Norway/2255/2013, both clade 3C -3 viruses, cover 3C-2 and 3C-3 quite well while maintaining adequate predicted coverage of clades 5, 6 and 3A. Furthermore A/Estonia/76614/2013 shows good predicted coverage of 3C-2, and 3C-3, as well as clades 3A, 5 and 6, while not covering clade 3C-1 as well. This lends credence to the prospect of a switch from the current clade 3C-1 vaccine candidates to a candidate from clade 3C-3 or 3C-2. Of the recently WHO approved manufacturing seeds, the 3C-3 viruses perform slightly better with A/Almaty/2958/2013 performing the best. We believe a 3C-3 vaccine candidate is possible as the next vaccine selection.

Table 13 Predicted Antigenic Coverage of Influenza A H3N2 Candidate Viruses in DASH

Antigen	InMedian	InLB	InUB	ObsIn	ObsOut	ObsUnk	DistIn	DistOut	Seqs
A/Cameroon/12V-5136/2012	85.6%	16.4%	100.0%	89.2%	0.4%	10.4%	45	2	1715
A/Estonia/76676/2013	77.2%	35.4%	100.0%	79.9%	0.4%	19.6%	42	2	1715
A/Estonia/76614/2013	75.6%	38.4%	89.2%	79.9%	0.4%	19.6%	42	2	1715
A/Norway/2255/2013	75.6%	35.1%	89.2%	77.1%	3.3%	19.6%	41	3	1715
A/Hong Kong/1036/2013	74.7%	37.1%	100.0%	76.6%	3.8%	19.6%	39	5	1715
A/Minsk/1262/2013	70.9%	31.5%	100.0%	77.1%	3.3%	19.6%	41	3	1715
A/Belgium/G1034/2013	70.6%	21.5%	85.9%	76.6%	3.8%	19.6%	39	5	1715
A/Cameroon/12V-5507/2012	70.1%	7.6%	100.0%	89.2%	0.4%	10.4%	45	2	1715
A/Latvia/1-32593/2013	70.1%	11.1%	85.7%	75.9%	3.0%	21.1%	42	6	1715
A/Slovenia/709/2013	70.0%	30.3%	94.1%	77.1%	0.6%	22.3%	45	3	1715
A/Almaty/2958/2013	60.0%	24.0%	75.9%	68.3%	3.8%	27.9%	43	5	1715
A/Victoria/361/2011 Cell	46.0%	38.7%	54.6%	66.7%	6.6%	26.7%	727	74	1715
A/Victoria/361/2011* Egg	45.8%	38.1%	51.9%	65.6%	8.2%	26.2%	727	74	1715

A/Texas/50/2012* Egg	16.7%	12.1%	22.1%	18.2%	51.2%	30.6%	398	403	1715
A/Texas/50/2012 Cell	10.8%	3.4%	94.1%	14.6%	3.5%	81.8%	38	6	1715
A/Perth/16/2009‡	5.8%	3.6%	8.7%	3.9%	69.3%	26.8%	134	650	1715

*Vaccine candidate

‡Previous vaccine candidate

Table 14 Analysis of Recent WHO Candidate Viruses.

Antigen	InMedian	InLB	InUB	ObsIn	ObsOut	ObsUnk	DistIn	DistOut	Seqs
A/Almaty/2958/2013	60.0%	24.0%	75.9%	68.3%	3.8%	27.9%	43	5	1715
A/New York/39/2012 Egg	42.7%	9.9%	85.3%	52.7%	0.5%	46.7%	25	3	1715
A/Serbia/NS-210/2013	42.0%	3.3%	79.4%	79.6%	3.9%	16.5%	31	6	1715
A/New York39/2012 Cell	38.0%	8.6%	85.3%	40.0%	3.4%	56.6%	24	4	1715
A/American Samoa/4768/2013 Egg	33.7%	18.2%	77.4%	43.4%	27.9%	28.6%	23	5	1715
A/American Samoa/4768/2013 Cell	25.2%	1.0%	48.9%	25.9%	4.5%	69.5%	18	10	1715

Results of simulations of antigenic coverage of circulating viruses are reported for each subtype and lineage in Tables 13 and 14. The table columns list antigen name followed by several performance statistics. InMedian means from 80 bootstrap replicates an average of X% of circulating viruses were predicted to be covered by the tested antigen. InLB and InUB are the lower and upper bounds of the bootstrap results for predicted coverage at a 95% confidence cutoff. ObsIn, ObsOut, and ObsUkn are the actual observed coverage predictions for circulating viruses when all the HI derived antigenic distance cohort measurements are used. DistIn and DistOut are the number of HI derived antigenic distance cohort measurements that were observed to fall in and out with respect to the test antigen at an antigenic distance of 2. Seqs is the weight-adjusted number of sequences determined to be circulating during the test period. These results were ordered by InMedian, followed by ObsIn, and finally DistIn.

H1N1 Analyses

H1N1 proportion tracking

Windowed proportion tracking analysis was performed on H1N1 sequences isolated between Jan 1, 2009 and December 1, 2013 using the HA1 domain. The analysis was conducted on the HA1 domain using a windowed partitioning scheme as described above. The current vaccine strain, A/California/7/2009, falls in the largest cluster of the final partition (2013_11), accounting for 59% of the total H1N1pdm sequences.

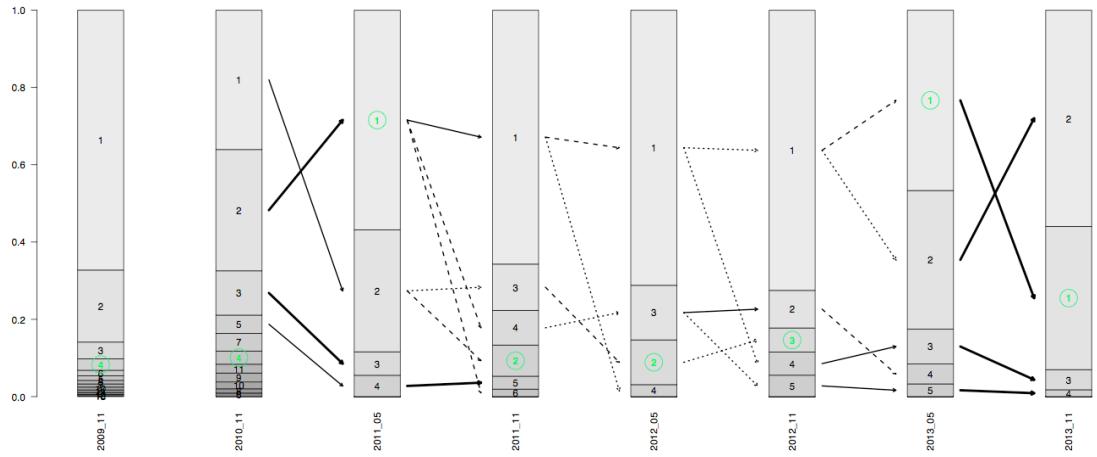


Figure 46 Proportional Sizes of H1N1 Clusters as Fraction of Total Sequences.

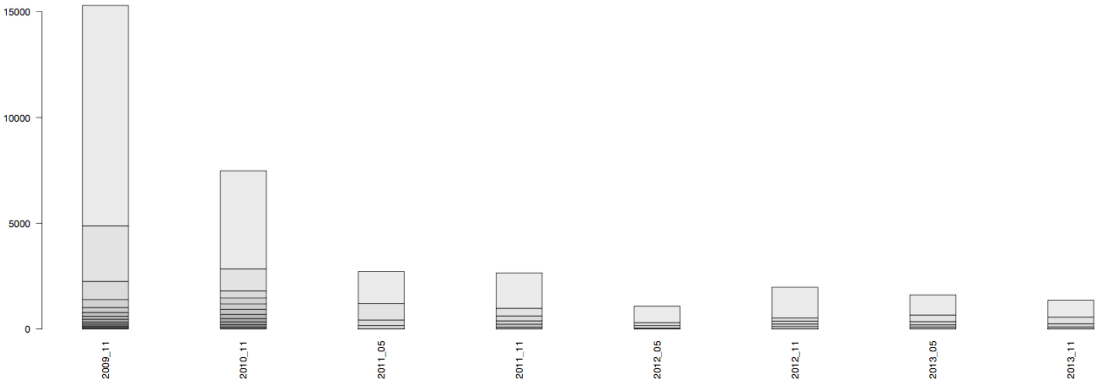


Figure 47 Sequence Counts Used for H1N1 Windowed Proportion Tracking.

DASH – MAR candidate selection

Analysis of DASH results for clusters without HI cohorts and candidate antigenic coverage prediction produced the following list of antigens that should be constructed as vaccine candidates or viruses to fill holes in HI antigen data.

PTP Cluster 2 candidates (Clade 6B):

A/England/358/2013

A/Bolivia/559/2013 (WHO candidate virus)

PTP Cluster 1 candidates (Clade 6C):

A/Estonia/74816/2013

A/Dominican Republic/7293/2013 (WHO candidate virus)

DASH gap filling viruses:

A/Puerto Rico/1/2013

A/Florida/43/2013

A/New Hampshire/4/2013

A/Minnesota/26/2012

A/Nizhny Novgorod/RII02/2013

It should be noted that in our analysis A/California/7/2009 is frequently surpassed in predicted coverage by a number of other viruses (only the top 4 candidates are shown in table 15). At an antigenic distance of 2, a vaccine containing the current recommendation as its H1N1 component is predicted to cover only 71.5% of circulating viruses. In contrast, an A/Estonia/74816/2013 containing vaccine is predicted to cover 100% of circulating viruses. A/England/358/2013 is predicted to cover 100% of circulating viruses as well. An analysis of the antigenic distance data

alone indicates that A/California/7/2009 is still within an antigenic distance of 2 of 96% of the viruses that it was tested against. While the performance of A/California/7/2009 has dipped in the DASH analysis, there is currently no evidence of a new antigenic group emerging to take its place. A closer look at the manufacturing seed A/Christchurch/16/2010, however shows poor performance against the currently circulating viruses. It is our opinion that A/Christchurch/16/2010 should be replaced, if it is currently being used in the formulation of vaccines, and replaced with a clade 6B virus such as A/Bolivia/559/2013 or A/England/358/2013.

Table 15 Predicted Antigenic Coverage of Influenza A Pandemic H1N1 Candidate Viruses in DASH

Antigen	InMedian	InLB	InUB	ObsIn	ObsOut	ObsUnk	DistIn	DistOut	Seqs
A/Estonia/74816/2013	100.0%	60.6%	100.0%	100.0%	0.0%	0.0%	56	0	1341
A/Luxembourg/80/2013	100.0%	59.4%	100.0%	100.0%	0.0%	0.0%	56	0	1341
A/England/358/2013	100.0%	31.8%	100.0%	100.0%	0.0%	0.0%	54	1	1341
A/Luxembourg/13/2013	100.0%	60.5%	100.0%	75.5%	0.1%	24.4%	55	1	1341
A/California/7/2009*	71.5%	54.7%	79.5%	85.2%	2.5%	12.3%	810	36	1341
A/Christchurch/16/2010	9.9%	6.7%	16.6%	7.7%	57.1%	35.2%	193	585	1341

*Vaccine Candidate

Table 16 Analysis of Recent WHO Candidate Viruses.

Antigen	InMedian	InLB	InUB	ObsIn	ObsOut	ObsUnk	DistIn	DistOut	Seqs
A/Bolivia/559/2013 Egg	70.1%	16.9%	100.0%	83.7%	0.1%	16.2%	16	1	1341
A/Dominican Republic/7293/2013 Cell	70.0%	10.2%	100.0%	83.7%	0.1%	16.2%	16	1	1341
A/Bolivia/559/2013 Cell	68.6%	15.5%	100.0%	83.7%	0.1%	16.2%	16	1	1341
A/Dominican Republic/7293/2013 Egg	68.0%	2.4%	100.0%	83.7%	0.1%	16.2%	16	1	1341

Influenza B/Yamagata Analyses

Influenza B/Yamagata Proportion Tracking

Proportion tracking analysis was conducted on all Influenza B/Yamagata lineage sequences isolated from January 1, 2001 until December 1, 2013. The analysis was conducted on the HA1 domain using a windowed partitioning scheme as described above. The current recommended vaccine seed, B/Massachusetts/2/2012 is seen in the largest cluster 1 of the current partition (2013_05) accounting for 86% of the total Influenza B Yamagata virus.

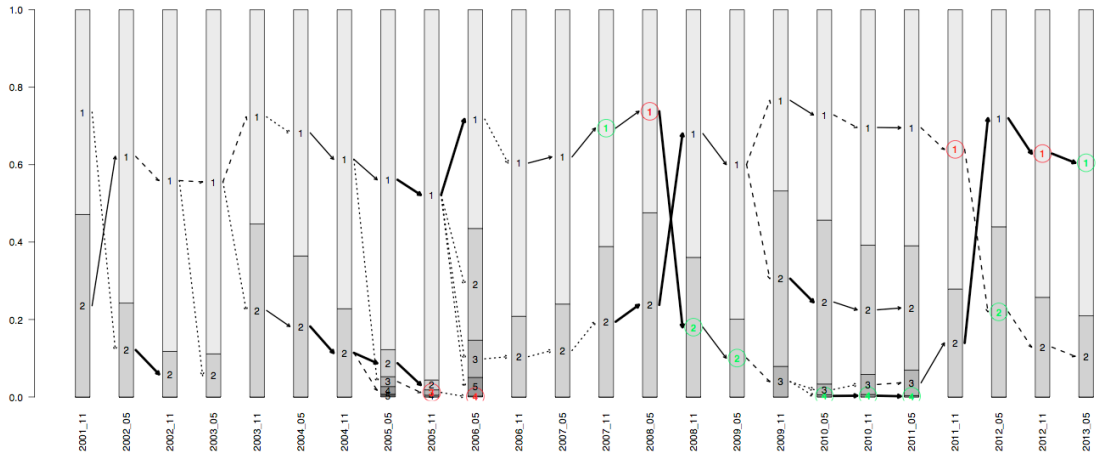


Figure 48 Proportional Sizes of Influenza B/Yamagata Clusters as Fraction of Total Sequences. Partition 2013_05 cluster 1 represents Yamagata Group 2 viruses.

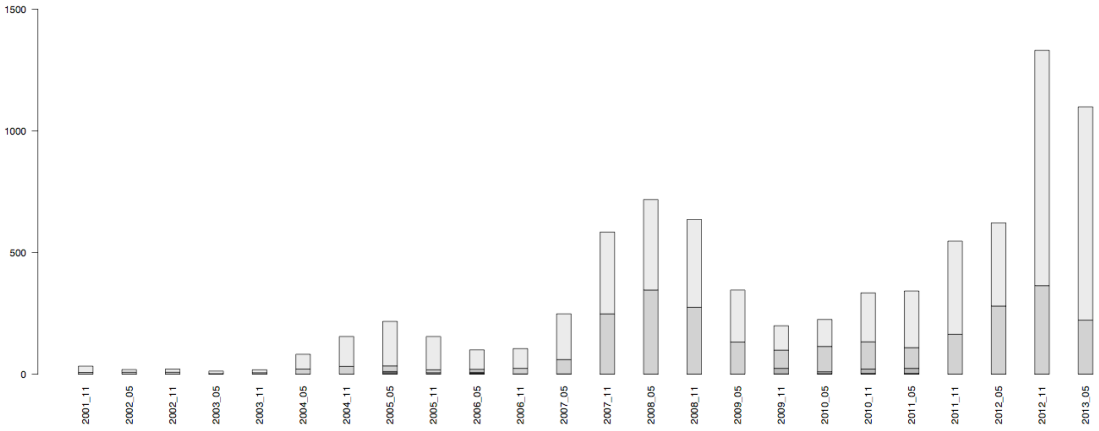


Figure 49 Sequence Counts Used for Influenza B/Yamagata Windowed Proportion Tracking.

DASH – MAR candidate selection

Analysis of DASH results for candidate antigenic coverage prediction produced the following list of antigens that should be constructed as vaccine candidates or viruses to fill holes in HI antigen data.

PTP cluster 1 (Group 2) candidates:

B/Massachusetts/2/2012*

PTP cluster 2 (Group 3) candidates:

B/Lithuania/6942/2013

DASH gap filling viruses:

None

It should be noted that group 2 of the Yamagata lineage shows the majority of new infections, and the current vaccine seed, B/Massachusetts/ 2/2012, performs quite well. This suggests that a new vaccine seed is probably not needed at this time, however the best high performing vaccine seed candidate from group 3 has been suggested above.

Table 17 Predicted Antigenic Coverage of Influenza B Yamagata Candidate Viruses in DASH

Antigen	InMedian	InLB	InUB	ObsIn	ObsOut	ObsUnk	DistIn	DistOut	Seqs
B/Massachusetts/2/2012 *	100.0%	8.0%	100.0%	100.0%	0.0%	0.0%	42	11	834
B/Novosibirsk/13/2013	92.9%	49.7%	100.0%	100.0%	0.0%	0.0%	59	2	834
B/Poland/8/2013	92.7%	58.3%	100.0%	100.0%	0.0%	0.0%	66	3	834
B/Omsk/35/2013	89.5%	42.0%	100.0%	100.0%	0.0%	0.0%	59	2	834
B/Ostrava/59/2013	87.7%	46.6%	100.0%	100.0%	0.0%	0.0%	58	3	834
B/Lithuania/6942/2013	86.9%	20.0%	100.0%	100.0%	0.0%	0.0%	60	1	834
B/Wisconsin/1/2010‡	35.2%	28.0%	40.5%	42.3%	40.4%	17.3%	824	597	834

*Vaccine candidate

‡Previous vaccine candidate

Influenza B/Victoria Analyses

Influenza B/Victoria Proportion Tracking

Proportion tracking analysis was conducted on all Influenza B/Victoria lineage sequences isolated from January 1, 2001 until December 1, 2013. The analysis was conducted on the HA1 domain using a windowed partitioning scheme as described above. The current recommended strain for the quadrivalent vaccine includes an additional Victoria lineage B/Brisbane/60/2008-like virus along with the recommended Yamagata lineage B/Massachusetts/2/2012-like virus. B/Brisbane/60/2008 is seen in the largest cluster 1 in the current partition (2013_05) accounting for 67% of the total Influenza B Yamagata virus circulating.

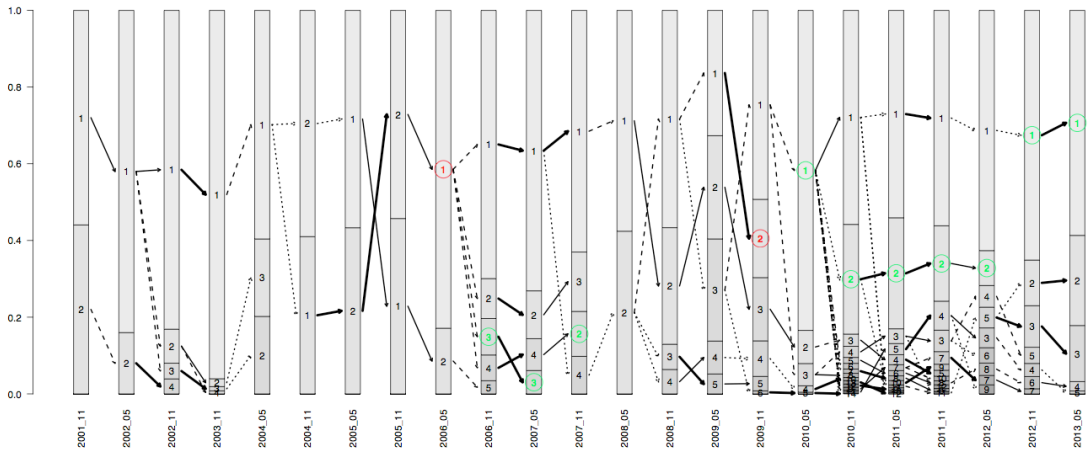


Figure 50 Proportional Sizes of Influenza B/Victoria Clusters as Fraction of Total Sequences.

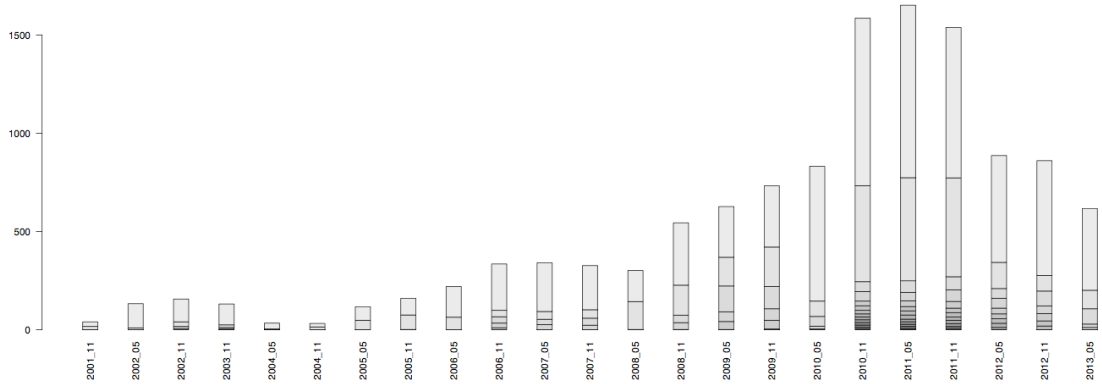


Figure 51 Sequence Counts Used for Influenza B/Victoria Windowed Proportion Tracking.

DASH – MAR candidate selection

Analysis of DASH results for candidate antigenic coverage prediction produced the following list of antigens that should be constructed as vaccine candidates or viruses to fill holes in HI antigen data.

PTP cluster 1 candidates:

B/Belgium/G886/2012

PTP cluster 2 candidates:

B/Texas/2/2013

PTP cluster 3 candidates:

B/ Formosa Province/V2367/2012

DASH gap filling viruses:

B/Iowa/4/2013

B/Sao Paulo/2-22035/2013

B/Indonesia/Nihrd-Buas704/2013

It should be noted that the current quadrivalent vaccine candidate for the Victoria lineage, B/Brisbane/60/2008, shows rather poor predicted coverage of currently circulating viruses. Better performing vaccine candidates, as listed above, exist for this lineage. Of these, a priority should be given to candidates from the larger PTP cluster 1, as the vast majority of new infections are members of this cluster. In PTP cluster 2, the recently announced WHO candidate B/Texas/2/2013 performs better than Brisbane/60 and appears to be egg stable. The overall prediction for Texas/2 was not great compared to other potential candidates. B/Texas/2/2013 was tested against a limited number of viruses, thus the supporting data for this analysis is not abundant.

Table 18 Predicted Antigenic Coverage of Influenza B Victoria Candidate Viruses in DASH

Antigen	InMedian	InLB	InUB	ObsIn	ObsOut	ObsUnk	DistIn	DistOut	Seqs
B/Formosa Province/V2367/2012	61.3%	36.8%	76.7%	70.9%	7.7%	21.5%	220	41	377
B/Belgium/G886/2012	45.2%	12.7%	100.0%	47.3%	1.1%	51.6%	30	5	377
B/Finland/310/2013	36.7%	12.4%	100.0%	47.3%	1.1%	51.6%	30	5	377
A/Texas/2/2013 Egg	32.3%	3.7%	72.8%	34.1%	1.3%	64.6%	33	2	377
B/Texas/2/2013 Cell	31.9%	19.5%	84.1%	32.5%	34.5%	33.0%	30	5	377
B/Dakar/18/2013	31.6%	2.7%	75.9%	32.5%	34.5%	33.0%	30	5	377
B/Dakar/10/2013	31.2%	2.1%	84.1%	32.2%	34.7%	33.0%	29	6	377
B/Brisbane/60/2008 *	25.1%	15.9%	34.9%	29.9%	44.7%	25.5%	156	182	377

*Vaccine candidate

Combined Influenza B Analyses

Influenza B Combined Proportion Tracking

Proportion tracking for combined Influenza B lineages was performed for sequences isolated from January 1, 2001 until December 1, 2013. The analysis was conducted on the HA1 domain using a windowed partitioning scheme. The combined proportion tracking for both Influenza B lineages shows that the Yamagata lineage

has continued to be the dominant lineage (cluster 1) in this current season accounting for 64% of the total Influenza B virus in circulation.

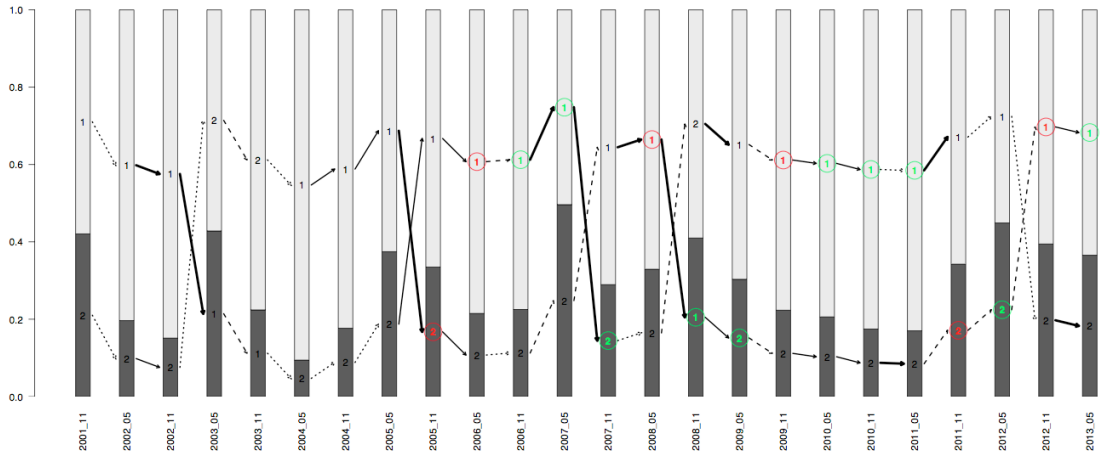


Figure 52 Proportional Sizes of Combined Influenza B Clusters as Fraction of Total Sequences. Partition 2013_05 cluster 1 are Yamagata lineage viruses, Partition 2013_05 cluster 2 are Victoria lineage viruses.

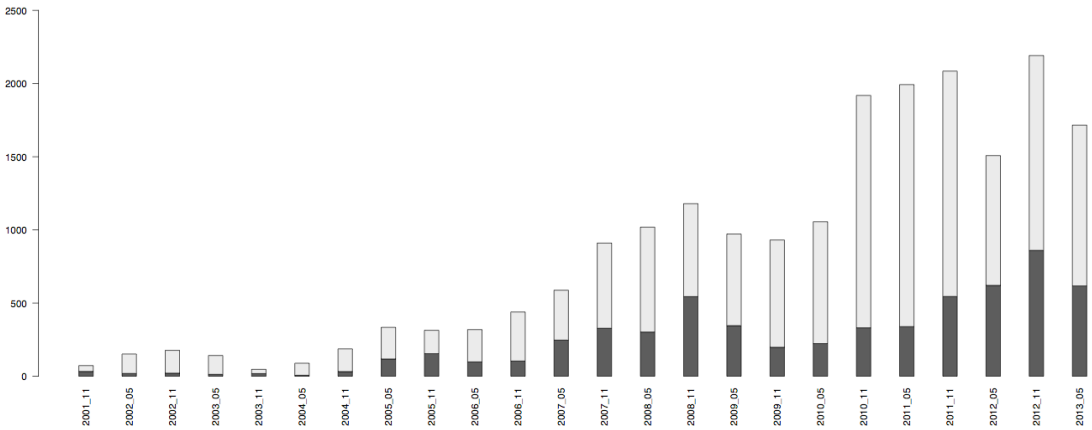


Figure 53 Sequence Counts Used for Influenza B Combined Windowed Proportion Tracking.

This provides evidence for continuing with a Yamagata lineage recommendation in the trivalent vaccine composition. However, since the Victoria lineage still accounts for 36% of the total virus in circulation, the recommendation to include a virus from the Victoria lineage along with the recommended B/Massachusetts/2/2012-like virus

from the Yamagata lineage in the quadrivalent vaccine should cover both of the lineages circulating in the population.

Appendix B. Glycosylation and dN/dS data supporting the RSV Study

Table 19 N-linked and O-linked Glycosylation Sites on the F Protein for 71 RSV Study Samples. Colored backgrounds reflect clade and/or genotype membership (green = GA5, purple = ON1, light blue = TN1, blue = TN2, red = BA.1, orange = BA.2).

lineage	accession	strain	244	27	70	116	120	126	clade	subclade	g_duplication	constellation_95
RSVA	KJ672479.1	RSVA/Homo_sapiens/USA/LA2_18/2013		NITE	NGTD	NYTL	NNTK	NVTL	2	2	0	15
RSVA	KJ672483.1	RSVA/Homo_sapiens/USA/LA2_19/2013		NITE	NGTD	NYTL	NNTK	NVTL	2	2	0	15
RSVA	KJ672474.1	RSVA/Homo_sapiens/USA/LA2_38/2012		NITE	NGTD	NYTL	NNTK	NVTL	2	2	0	15
RSVA	KJ672447.1	RSVA/Homo_sapiens/USA/LA2_67/2013		NITE	NGTD	NYTL	NNTK	NVTL	2	2	0	15
RSVA	KJ672462.1	RSVA/Homo_sapiens/USA/LA2_84/2013		NITE	NGTD	NYTL	NNTK	NVTL	2	2	0	15
RSVA	KJ672464.1	RSVA/Homo_sapiens/USA/LA2_04/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672468.1	RSVA/Homo_sapiens/USA/LA2_07/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672466.1	RSVA/Homo_sapiens/USA/LA2_08/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	45567-MAIN	RSVA/Homo_sapiens/USA/LA2_10/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672457.1	RSVA/Homo_sapiens/USA/LA2_100/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672451.1	RSVA/Homo_sapiens/USA/LA2_103/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672461.1	RSVA/Homo_sapiens/USA/LA2_105/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672424.1	RSVA/Homo_sapiens/USA/LA2_106/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672444.1	RSVA/Homo_sapiens/USA/LA2_11/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672465.1	RSVA/Homo_sapiens/USA/LA2_13/2012	+	NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672472.1	RSVA/Homo_sapiens/USA/LA2_15/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672475.1	RSVA/Homo_sapiens/USA/LA2_17/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672440.1	RSVA/Homo_sapiens/USA/LA2_22/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672449.1	RSVA/Homo_sapiens/USA/LA2_28/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672470.1	RSVA/Homo_sapiens/USA/LA2_34/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672469.1	RSVA/Homo_sapiens/USA/LA2_44/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	45606-MAIN	RSVA/Homo_sapiens/USA/LA2_49/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672428.1	RSVA/Homo_sapiens/USA/LA2_55/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	45613-MAIN	RSVA/Homo_sapiens/USA/LA2_56/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672480.1	RSVA/Homo_sapiens/USA/LA2_62/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672429.1	RSVA/Homo_sapiens/USA/LA2_69/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672432.1	RSVA/Homo_sapiens/USA/LA2_72/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672454.1	RSVA/Homo_sapiens/USA/LA2_73/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672448.1	RSVA/Homo_sapiens/USA/LA2_74/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672458.1	RSVA/Homo_sapiens/USA/LA2_77/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	45640-MAIN	RSVA/Homo_sapiens/USA/LA2_83/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672471.1	RSVA/Homo_sapiens/USA/LA2_85/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672441.1	RSVA/Homo_sapiens/USA/LA2_87/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	45646-MAIN	RSVA/Homo_sapiens/USA/LA2_90/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672452.1	RSVA/Homo_sapiens/USA/LA2_91/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	45650-MAIN	RSVA/Homo_sapiens/USA/LA2_94/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672433.1	RSVA/Homo_sapiens/USA/LA2_95/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672467.1	RSVA/Homo_sapiens/USA/LA2_97/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	45654-MAIN	RSVA/Homo_sapiens/USA/LA2_98/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672437.1	RSVA/Homo_sapiens/USA/LA2_99/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1a	1	1
RSVA	KJ672484.1	RSVA/Homo_sapiens/USA/LA2_01/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	KJ672439.1	RSVA/Homo_sapiens/USA/LA2_02/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	KJ672455.1	RSVA/Homo_sapiens/USA/LA2_05/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	KJ672427.1	RSVA/Homo_sapiens/USA/LA2_09/2012	+	NITE	NGTD	NYTL	NNTK		1	1b	0	3
RSVA	KJ672431.1	RSVA/Homo_sapiens/USA/LA2_14/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	KJ672443.1	RSVA/Homo_sapiens/USA/LA2_21/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	KJ672456.1	RSVA/Homo_sapiens/USA/LA2_26/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3

lineage	accession	strain	244	27	70	116	120	126	clade	subclade	g_duplication	constellation_95
RSVA	KJ672478.1	RSVA/Homo_sapiens/USA/LA2_29/2013	+	NITE	NGTD	NYTL		NVTL	1	1b	0	3
RSVA	KJ672463.1	RSVA/Homo_sapiens/USA/LA2_31/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	KJ672477.1	RSVA/Homo_sapiens/USA/LA2_33/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	KJ672426.1	RSVA/Homo_sapiens/USA/LA2_36/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	KJ672435.1	RSVA/Homo_sapiens/USA/LA2_37/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	KJ672460.1	RSVA/Homo_sapiens/USA/LA2_40/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	KJ672482.1	RSVA/Homo_sapiens/USA/LA2_41/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	KJ672450.1	RSVA/Homo_sapiens/USA/LA2_46/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	KJ672453.1	RSVA/Homo_sapiens/USA/LA2_48/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	KJ672434.1	RSVA/Homo_sapiens/USA/LA2_53/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	45648-MAIN	RSVA/Homo_sapiens/USA/LA2_92/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	0	3
RSVA	KJ672446.1	RSVA/Homo_sapiens/USA/LA2_27/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	1	1
RSVA	45596-MAIN	RSVA/Homo_sapiens/USA/LA2_39/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	1	1
RSVA	KJ672442.1	RSVA/Homo_sapiens/USA/LA2_45/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	1	1
RSVA	45635-MAIN	RSVA/Homo_sapiens/USA/LA2_78/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1b	1	1
RSVA	KJ672436.1	RSVA/Homo_sapiens/USA/LA2_30/2012		NITE	NGTD	NYTL	NNTK	NVTL	1	1c	0	3
RSVA	KJ672459.1	RSVA/Homo_sapiens/USA/LA2_60/2013		NITE	NGTD	NYTL	NNTK	NVTL	1	1c	0	3
RSVB	KJ672476.1	RSVB/Homo_sapiens/USA/LA2_42/2013		NITE	NGTD	NYTI	NTTK	NVSI	3	3a	1	22
RSVB	45573-MAIN	RSVB/Homo_sapiens/USA/LA2_16/2012		NITE	NGTD	NYTI	NTTK	NVSI	3	3b	1	19
RSVB	KJ672473.1	RSVB/Homo_sapiens/USA/LA2_24/2012		NITE	NGTD	NYTI	NTTK	NVSI	3	3b	1	19
RSVB	KJ672425.1	RSVB/Homo_sapiens/USA/LA2_25/2013		NITE	NGTD	NYTI		NVSI	3	3b	1	19
RSVB	KJ672438.1	RSVB/Homo_sapiens/USA/LA2_50/2013		NITE	NGTD	NYTI	NTTK	NVSI	3	3b	1	19
RSVB	KJ672430.1	RSVB/Homo_sapiens/USA/LA2_51/2013		NITE	NGTD	NYTI	NTTK	NVSI	3	3b	1	19
RSVB	KJ672481.1	RSVB/Homo_sapiens/USA/LA2_82/2013		NITE	NGTD	NYTI	NTTK	NVSI	3	3b	1	19

Table 20 N-linked and O-linked Glycosylation Sites on the G Protein for 71 RSV Study Samples. Colored backgrounds reflect clade and/or genotype membership (green = GA5, purple = ON1, light blue = TN1, blue = TN2, red = BA.1, orange = BA.2).

lineage	accession	strain_name	64	70	72	73	75	78	80	81	83	86	87	88	89	92	95	100	101	102	105	106	107
RSVA	KJ672479.1	RSVA/Homo_sapiens/USA/LA2_18/2013	+	+	+	+			+	+		+	+		+	+		+			+		+
RSVA	KJ672483.1	RSVA/Homo_sapiens/USA/LA2_19/2013		+	+	+			+	+		+	+		+	+		+			+		+
RSVA	KJ672474.1	RSVA/Homo_sapiens/USA/LA2_38/2012	+	+	+	+			+	+		+	+		+	+		+			+		+
RSVA	KJ672447.1	RSVA/Homo_sapiens/USA/LA2_67/2013	+	+	+	+			+	+		+	+		+	+		+			+		+
RSVA	KJ672462.1	RSVA/Homo_sapiens/USA/LA2_84/2013	+	+	+	+			+	+		+	+		+	+		+			+		+
RSVA	KJ672464.1	RSVA/Homo_sapiens/USA/LA2_04/2012		+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672468.1	RSVA/Homo_sapiens/USA/LA2_07/2012	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672466.1	RSVA/Homo_sapiens/USA/LA2_08/2012		+	+	+			+			+	+		+	+		+		+	+		+
RSVA	45567-MAIN	RSVA/Homo_sapiens/USA/LA2_10/2012	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672457.1	RSVA/Homo_sapiens/USA/LA2_100/2013	+	+	+				+			+	+		+	+		+		+	+		+
RSVA	KJ672451.1	RSVA/Homo_sapiens/USA/LA2_103/2012		+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672461.1	RSVA/Homo_sapiens/USA/LA2_105/2012		+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672424.1	RSVA/Homo_sapiens/USA/LA2_106/2012		+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672444.1	RSVA/Homo_sapiens/USA/LA2_11/2012	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672465.1	RSVA/Homo_sapiens/USA/LA2_13/2012		+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672472.1	RSVA/Homo_sapiens/USA/LA2_15/2012	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672475.1	RSVA/Homo_sapiens/USA/LA2_17/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672440.1	RSVA/Homo_sapiens/USA/LA2_22/2012	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672449.1	RSVA/Homo_sapiens/USA/LA2_28/2012	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672470.1	RSVA/Homo_sapiens/USA/LA2_34/2012		+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672469.1	RSVA/Homo_sapiens/USA/LA2_44/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	45606-MAIN	RSVA/Homo_sapiens/USA/LA2_49/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672428.1	RSVA/Homo_sapiens/USA/LA2_55/2013	+		+	+			+			+	+		+	+		+		+	+		+
RSVA	45613-MAIN	RSVA/Homo_sapiens/USA/LA2_56/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672480.1	RSVA/Homo_sapiens/USA/LA2_62/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672429.1	RSVA/Homo_sapiens/USA/LA2_69/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672432.1	RSVA/Homo_sapiens/USA/LA2_72/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672454.1	RSVA/Homo_sapiens/USA/LA2_73/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672448.1	RSVA/Homo_sapiens/USA/LA2_74/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672458.1	RSVA/Homo_sapiens/USA/LA2_77/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	45640-MAIN	RSVA/Homo_sapiens/USA/LA2_83/2013	+	+	+				+			+	+		+	+		+		+	+		+
RSVA	KJ672471.1	RSVA/Homo_sapiens/USA/LA2_85/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672441.1	RSVA/Homo_sapiens/USA/LA2_87/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	45646-MAIN	RSVA/Homo_sapiens/USA/LA2_90/2012		+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672452.1	RSVA/Homo_sapiens/USA/LA2_91/2013		+	+	+			+			+	+		+	+		+		+	+		+
RSVA	45650-MAIN	RSVA/Homo_sapiens/USA/LA2_94/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672433.1	RSVA/Homo_sapiens/USA/LA2_95/2013	+	+	+	+			+			+	+		+	+		+	+	+	+		+
RSVA	KJ672467.1	RSVA/Homo_sapiens/USA/LA2_97/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	45654-MAIN	RSVA/Homo_sapiens/USA/LA2_98/2013	+	+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672437.1	RSVA/Homo_sapiens/USA/LA2_99/2013	+	+	+				+			+	+		+	+		+		+	+		+
RSVA	KJ672484.1	RSVA/Homo_sapiens/USA/LA2_01/2012	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672439.1	RSVA/Homo_sapiens/USA/LA2_02/2012	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672455.1	RSVA/Homo_sapiens/USA/LA2_05/2012	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672427.1	RSVA/Homo_sapiens/USA/LA2_09/2012	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672431.1	RSVA/Homo_sapiens/USA/LA2_14/2012	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672443.1	RSVA/Homo_sapiens/USA/LA2_21/2013	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672456.1	RSVA/Homo_sapiens/USA/LA2_26/2012	+	+	+	+						+	+		+	+		+		+	+		+

lineage	accession	strain_name	64	70	72	73	75	78	80	81	83	86	87	88	89	92	95	100	101	102	105	106	107
RSVA	KJ672478.1	RSVA/Homo_sapiens/USA/LA2_29/2013	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672463.1	RSVA/Homo_sapiens/USA/LA2_31/2013		+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672477.1	RSVA/Homo_sapiens/USA/LA2_33/2012	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672426.1	RSVA/Homo_sapiens/USA/LA2_36/2013	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672435.1	RSVA/Homo_sapiens/USA/LA2_37/2012	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672460.1	RSVA/Homo_sapiens/USA/LA2_40/2013	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672482.1	RSVA/Homo_sapiens/USA/LA2_41/2012	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672450.1	RSVA/Homo_sapiens/USA/LA2_46/2013	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672453.1	RSVA/Homo_sapiens/USA/LA2_48/2013	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672434.1	RSVA/Homo_sapiens/USA/LA2_53/2013	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	45648-MAIN	RSVA/Homo_sapiens/USA/LA2_92/2013	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672446.1	RSVA/Homo_sapiens/USA/LA2_27/2012	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	45596-MAIN	RSVA/Homo_sapiens/USA/LA2_39/2013	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672442.1	RSVA/Homo_sapiens/USA/LA2_45/2013	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	45635-MAIN	RSVA/Homo_sapiens/USA/LA2_78/2013	+	+	+	+						+	+		+	+		+		+	+		+
RSVA	KJ672436.1	RSVA/Homo_sapiens/USA/LA2_30/2012		+	+	+			+			+	+		+	+		+		+	+		+
RSVA	KJ672459.1	RSVA/Homo_sapiens/USA/LA2_60/2013		+	+	+			+			+	+		+	+		+		+	+		+
RSVB	KJ672476.1	RSVB/Homo_sapiens/USA/LA2_42/2013	+	+	+	+	+	+			+			+	+	+	+	+		+	+	+	+
RSVB	45573-MAIN	RSVB/Homo_sapiens/USA/LA2_16/2012		+	+	+	+	+			+		+	+	+	+	+	+		+		+	+
RSVB	KJ672473.1	RSVB/Homo_sapiens/USA/LA2_24/2012		+	+	+	+	+			+		+	+	+	+	+	+		+		+	
RSVB	KJ672425.1	RSVB/Homo_sapiens/USA/LA2_25/2013		+	+	+	+	+			+		+	+	+	+	+	+		+		+	+
RSVB	KJ672438.1	RSVB/Homo_sapiens/USA/LA2_50/2013	+	+	+	+	+				+		+	+	+	+	+	+		+		+	+
RSVB	KJ672430.1	RSVB/Homo_sapiens/USA/LA2_51/2013		+	+	+	+	+			+		+	+	+	+	+	+		+		+	+
RSVB	KJ672481.1	RSVB/Homo_sapiens/USA/LA2_82/2013		+	+	+	+	+			+		+	+	+	+	+	+		+		+	+

strain_name	108	109	111	112	113	115	117	118	119	121	122	124	125	126	128	129	130	133	136	137	138	139	141	144	146	147
RSVA/Homo_sapiens/USA/LA2_18/2013		+	+	+				+	+	+					+	+	+	+	+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_19/2013		+	+	+				+	+	+					+	+	+	+	+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_38/2012		+	+	+				+	+	+					+	+	+	+	+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_67/2013		+	+	+				+	+	+					+	+	+	+	+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_84/2013		+	+	+				+	+	+					+	+	+	+	+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_04/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_07/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_08/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_10/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_100/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_103/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_105/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_106/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_11/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_13/2012	+	+	+	+	+		+	+	+	+		+	+			+	+		+	+		+		+		+
RSVA/Homo_sapiens/USA/LA2_15/2012	+	+	+	+	+		+	+	+	+		+	+			+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_17/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_22/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_28/2012	+	+	+	+	+		+	+	+	+			+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_34/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_44/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_49/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_55/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_56/2013	+	+	+	+	+		+	+	+	+		+	+			+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_62/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_69/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_72/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_73/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_74/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_77/2013	+	+	+	+	+		+	+	+	+		+	+			+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_83/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_85/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+	+	+
RSVA/Homo_sapiens/USA/LA2_87/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_90/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_91/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_94/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_95/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_97/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_98/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_99/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_01/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_02/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_05/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_09/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_14/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_21/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_26/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+

strain_name	108	109	111	112	113	115	117	118	119	121	122	124	125	126	128	129	130	133	136	137	138	139	141	144	146	147
RSVA/Homo_sapiens/USA/LA2_29/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_31/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_33/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_36/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_37/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_40/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_41/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_46/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_48/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_53/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_92/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_27/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_39/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_45/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_78/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+		+
RSVA/Homo_sapiens/USA/LA2_30/2012	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+	+	+
RSVA/Homo_sapiens/USA/LA2_60/2013	+	+	+	+	+		+	+	+	+		+	+		+	+	+		+	+	+	+		+	+	+
RSVB/Homo_sapiens/USA/LA2_42/2013	+				+	+	+		+		+	+		+		+	+	+		+	+	+	+			+
RSVB/Homo_sapiens/USA/LA2_16/2012	+				+	+	+		+		+	+		+		+	+	+		+	+	+	+			+
RSVB/Homo_sapiens/USA/LA2_24/2012	+				+	+	+		+		+	+		+			+	+		+	+		+			+
RSVB/Homo_sapiens/USA/LA2_25/2013	+				+	+	+		+		+	+		+		+	+	+		+	+	+	+			+
RSVB/Homo_sapiens/USA/LA2_50/2013	+				+	+	+		+		+	+		+		+	+	+		+	+	+	+			+
RSVB/Homo_sapiens/USA/LA2_51/2013	+				+	+	+		+		+	+		+		+	+	+		+	+	+	+			+
RSVB/Homo_sapiens/USA/LA2_82/2013	+				+	+	+		+		+	+		+		+	+	+		+	+	+	+			+

strain_name	148	172	174	177	181	182	186	189	191	197	198	199	200	201	203	204	206	207	208	209	210	211	216	219	220	225
RSVA/Homo_sapiens/USA/LA2_18/2013	+		+	+	+				+		+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_19/2013	+		+	+	+				+		+	+	+		+			+			+	+		+	+	+
RSVA/Homo_sapiens/USA/LA2_38/2012	+		+	+	+				+		+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_67/2013	+		+	+	+				+		+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_84/2013	+		+	+	+				+		+	+	+		+			+			+	+		+	+	+
RSVA/Homo_sapiens/USA/LA2_04/2012	+		+	+	+						+	+	+		+			+			+			+	+	
RSVA/Homo_sapiens/USA/LA2_07/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_08/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_10/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_100/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_103/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_105/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_106/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_11/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_13/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_15/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_17/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_22/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_28/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_34/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_44/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_49/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_55/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_56/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_62/2013	+		+	+	+						+	+	+		+				+					+	+	
RSVA/Homo_sapiens/USA/LA2_69/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_72/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_73/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_74/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_77/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_83/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_85/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_87/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_90/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_91/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_94/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_95/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_97/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_98/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_99/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_01/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_02/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_05/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_09/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_14/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_21/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_26/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	

strain_name	148	172	174	177	181	182	186	189	191	197	198	199	200	201	203	204	206	207	208	209	210	211	216	219	220	225
RSVA/Homo_sapiens/USA/LA2_29/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_31/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_33/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_36/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_37/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_40/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_41/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_46/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_48/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_53/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_92/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_27/2012	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_39/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_45/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_78/2013	+		+	+	+						+	+	+		+			+			+	+		+	+	
RSVA/Homo_sapiens/USA/LA2_30/2012	+		+	+	+						+	+	+		+			+			+	+			+	
RSVA/Homo_sapiens/USA/LA2_60/2013	+		+	+	+						+	+	+		+			+			+	+			+	
RSVB/Homo_sapiens/USA/LA2_42/2013	+	+				+	+	+		+				+			+		+	+			+		+	+
RSVB/Homo_sapiens/USA/LA2_16/2012	+	+				+	+	+		+				+		+	+		+	+			+		+	+
RSVB/Homo_sapiens/USA/LA2_24/2012	+	+				+	+	+		+				+		+	+		+	+			+		+	+
RSVB/Homo_sapiens/USA/LA2_25/2013	+	+				+	+	+		+				+		+	+		+	+			+		+	+
RSVB/Homo_sapiens/USA/LA2_50/2013	+					+	+	+		+				+		+	+		+	+			+		+	+
RSVB/Homo_sapiens/USA/LA2_51/2013	+	+				+	+	+		+				+		+	+		+	+			+		+	+
RSVB/Homo_sapiens/USA/LA2_82/2013	+	+				+	+	+		+				+		+	+		+	+			+		+	+

strain_name	226	227	228	230	231	234	235	236	237	238	239	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255
RSVA/Homo_sapiens/USA/LA2_18/2013		+	+		+		+	+			+		+			+	+			+		+	+	+		
RSVA/Homo_sapiens/USA/LA2_19/2013		+	+		+		+	+			+					+	+			+		+	+	+		
RSVA/Homo_sapiens/USA/LA2_38/2012		+	+		+		+	+			+		+			+	+			+		+	+	+		
RSVA/Homo_sapiens/USA/LA2_67/2013		+	+		+		+	+			+		+			+	+			+		+	+	+		
RSVA/Homo_sapiens/USA/LA2_84/2013		+	+		+		+	+			+					+	+			+		+	+	+		
RSVA/Homo_sapiens/USA/LA2_04/2012		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_07/2012		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_08/2012		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_10/2012		+	+		+		+			+	+					+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_100/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_103/2012		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_105/2012		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_106/2012		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_11/2012		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_13/2012		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_15/2012		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_17/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_22/2012		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_28/2012		+	+		+		+			+	+	+				+	+				+		+			
RSVA/Homo_sapiens/USA/LA2_34/2012		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_44/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_49/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_55/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_56/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_62/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_69/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_72/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_73/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_74/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_77/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_83/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_85/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_87/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_90/2012		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_91/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_94/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_95/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_97/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_98/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_99/2013		+	+		+		+			+	+	+				+	+			+	+		+			
RSVA/Homo_sapiens/USA/LA2_01/2012		+	+		+		+			+	+	+					+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_02/2012		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_05/2012		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_09/2012		+	+		+		+			+	+	+					+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_14/2012		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_21/2013		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_26/2012		+	+		+		+			+	+	+				+	+			+	+		+	+		

strain_name	226	227	228	230	231	234	235	236	237	238	239	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255
RSVA/Homo_sapiens/USA/LA2_29/2013		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_31/2013		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_33/2012		+	+		+		+			+	+	+					+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_36/2013		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_37/2012		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_40/2013		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_41/2012		+	+		+		+			+	+	+					+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_46/2013		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_48/2013		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_53/2013		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_92/2013		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_27/2012		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_39/2013		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_45/2013		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_78/2013		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_30/2012		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVA/Homo_sapiens/USA/LA2_60/2013		+	+		+		+			+	+	+				+	+			+	+		+	+		
RSVB/Homo_sapiens/USA/LA2_42/2013	+			+		+			+	+			+	+	+			+	+				+	+	+	+
RSVB/Homo_sapiens/USA/LA2_16/2012	+	+		+		+			+	+			+	+	+			+	+				+	+	+	+
RSVB/Homo_sapiens/USA/LA2_24/2012	+	+		+		+			+	+			+	+	+			+	+				+	+	+	+
RSVB/Homo_sapiens/USA/LA2_25/2013	+	+		+		+			+	+			+	+	+			+	+				+	+	+	+
RSVB/Homo_sapiens/USA/LA2_50/2013	+	+		+		+			+	+			+	+	+			+	+				+	+	+	+
RSVB/Homo_sapiens/USA/LA2_51/2013	+	+		+		+			+	+			+	+	+			+	+				+	+	+	+
RSVB/Homo_sapiens/USA/LA2_82/2013	+	+		+		+			+	+			+	+	+			+	+				+	+	+	+

strain_name	258	259	260	262	263	264	265	267	268	269	270	272	273	274	275	277	278	281	282	283	286	287	288	289	291	292
RSVA/Homo_sapiens/USA/LA2_18/2013		+	+			+		+	+	+	+			+	+	+		+	+	+						
RSVA/Homo_sapiens/USA/LA2_19/2013		+	+			+		+	+	+	+			+	+	+		+	+	+						
RSVA/Homo_sapiens/USA/LA2_38/2012		+	+			+		+		+	+			+	+	+		+	+	+						
RSVA/Homo_sapiens/USA/LA2_67/2013		+	+			+		+	+	+	+			+	+	+		+	+	+						
RSVA/Homo_sapiens/USA/LA2_84/2013		+	+			+		+	+	+	+			+	+	+		+	+	+						
RSVA/Homo_sapiens/USA/LA2_04/2012		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_07/2012		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_08/2012		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_10/2012		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_100/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_103/2012		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_105/2012		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_106/2012		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_11/2012		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_13/2012		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_15/2012		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_17/2013		+	+					+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_22/2012		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_28/2012		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_34/2012		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_44/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_49/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_55/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_56/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_62/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_69/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_72/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_73/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_74/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_77/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_83/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_85/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_87/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_90/2012		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_91/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_94/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_95/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_97/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_98/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_99/2013		+	+			+		+	+	+	+				+	+		+	+	+			+		+	+
RSVA/Homo_sapiens/USA/LA2_01/2012		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_02/2012		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_05/2012		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_09/2012		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_14/2012		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_21/2013		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_26/2012		+	+			+		+	+	+	+				+	+		+	+							

strain_name	258	259	260	262	263	264	265	267	268	269	270	272	273	274	275	277	278	281	282	283	286	287	288	289	291	292
RSVA/Homo_sapiens/USA/LA2_29/2013		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_31/2013		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_33/2012		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_36/2013		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_37/2012		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_40/2013		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_41/2012		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_46/2013		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_48/2013		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_53/2013		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_92/2013		+	+			+		+	+	+	+				+	+		+	+							
RSVA/Homo_sapiens/USA/LA2_27/2012		+	+			+		+	+	+	+				+	+		+	+				+		+	+
RSVA/Homo_sapiens/USA/LA2_39/2013		+	+			+		+	+	+	+				+	+		+	+				+		+	+
RSVA/Homo_sapiens/USA/LA2_45/2013		+	+			+		+	+	+	+				+	+		+	+				+		+	+
RSVA/Homo_sapiens/USA/LA2_78/2013		+	+			+		+	+	+	+				+	+		+	+				+		+	+
RSVA/Homo_sapiens/USA/LA2_30/2012		+	+			+		+	+	+	+				+	+		+	+	+						
RSVA/Homo_sapiens/USA/LA2_60/2013		+	+			+		+	+	+	+				+	+		+	+	+						
RSVB/Homo_sapiens/USA/LA2_42/2013	+			+	+	+	+	+				+	+	+	+		+			+	+	+	+			+
RSVB/Homo_sapiens/USA/LA2_16/2012	+			+	+	+	+	+				+	+	+	+		+			+	+	+	+			+
RSVB/Homo_sapiens/USA/LA2_24/2012	+			+	+	+	+	+				+	+	+	+		+			+	+	+	+			+
RSVB/Homo_sapiens/USA/LA2_25/2013	+			+	+	+	+	+				+	+	+	+		+			+	+	+	+			+
RSVB/Homo_sapiens/USA/LA2_50/2013	+			+	+	+	+	+				+	+	+	+		+			+	+	+	+			+
RSVB/Homo_sapiens/USA/LA2_51/2013	+			+	+	+	+	+				+	+	+	+		+			+	+	+	+			+
RSVB/Homo_sapiens/USA/LA2_82/2013	+			+	+	+	+	+				+	+	+	+		+			+	+	+	+			+

strain_name	293	294	295	296	298	299	300	301	302	305	306	307	309	311	313	315	316	317
RSVA/Homo_sapiens/USA/LA2_18/2013														+		+		+
RSVA/Homo_sapiens/USA/LA2_19/2013														+		+		+
RSVA/Homo_sapiens/USA/LA2_38/2012														+		+		+
RSVA/Homo_sapiens/USA/LA2_67/2013														+		+		+
RSVA/Homo_sapiens/USA/LA2_84/2013														+		+		+
RSVA/Homo_sapiens/USA/LA2_04/2012	+					+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_07/2012	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_08/2012	+					+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_10/2012	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_100/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_103/2012	+					+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_105/2012	+					+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_106/2012	+					+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_11/2012	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_13/2012	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_15/2012	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_17/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_22/2012	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_28/2012	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_34/2012	+					+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_44/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_49/2013	+					+				+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_55/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_56/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_62/2013	+					+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_69/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_72/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_73/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_74/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_77/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_83/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_85/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_87/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_90/2012	+					+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_91/2013	+					+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_94/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_95/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_97/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_98/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_99/2013	+	+				+		+		+	+	+		+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_01/2012														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_02/2012														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_05/2012														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_09/2012														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_14/2012														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_21/2013														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_26/2012														+	+	+	+	+

strain_name	81	86	103	135	209	228	237	250	251	294	308	318	clade	subclade
RSVA/Homo_sapiens/USA/LA2_18/2013			NLSG	NTTT			NITK	NSTT				NITD	2	2
RSVA/Homo_sapiens/USA/LA2_19/2013			NLSG	NTTT			NITK	NSTT				NITD	2	2
RSVA/Homo_sapiens/USA/LA2_38/2012			NLSG	NTTT			NITK	NSTT				NITD	2	2
RSVA/Homo_sapiens/USA/LA2_67/2013			NLSG	NTTT			NITK	NSTT				NITD	2	2
RSVA/Homo_sapiens/USA/LA2_84/2013			NLSG	NTTT			NITK	NSTT				NITD	2	2
RSVA/Homo_sapiens/USA/LA2_04/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_07/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_08/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_10/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_100/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_103/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_105/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_106/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_11/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_13/2012			NLSG	NTTI			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_15/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_17/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_22/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_28/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_34/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_44/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_49/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_55/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_56/2013			NLSG				NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_62/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_69/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_72/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_73/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_74/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_77/2013			NLSG				NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_83/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_85/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_87/2013			NLSG	NTTT			NTTK					NITK	1	1a
RSVA/Homo_sapiens/USA/LA2_90/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_91/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_94/2013			NLSG	NTTT	NTTK		NTTR						1	1a
RSVA/Homo_sapiens/USA/LA2_95/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_97/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_98/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_99/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_01/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_02/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_05/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_09/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_14/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_21/2013			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_26/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b

strain_name	293	294	295	296	298	299	300	301	302	305	306	307	309	311	313	315	316	317
RSVA/Homo_sapiens/USA/LA2_29/2013														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_31/2013														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_33/2012														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_36/2013														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_37/2012														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_40/2013														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_41/2012														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_46/2013														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_48/2013														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_53/2013														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_92/2013														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_27/2012	+	+				+		+		+	+			+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_39/2013	+	+				+		+		+	+			+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_45/2013	+	+				+		+		+	+			+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_78/2013	+	+				+		+		+	+			+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_30/2012														+	+	+	+	+
RSVA/Homo_sapiens/USA/LA2_60/2013														+	+	+	+	+
RSVB/Homo_sapiens/USA/LA2_42/2013			+	+	+				+	+	+	+						
RSVB/Homo_sapiens/USA/LA2_16/2012			+	+	+		+		+	+	+	+	+					
RSVB/Homo_sapiens/USA/LA2_24/2012			+	+	+		+		+	+	+	+	+					
RSVB/Homo_sapiens/USA/LA2_25/2013			+	+	+		+		+	+	+	+	+					
RSVB/Homo_sapiens/USA/LA2_50/2013			+	+	+		+		+	+	+	+	+					
RSVB/Homo_sapiens/USA/LA2_51/2013			+	+	+		+		+	+	+	+	+					
RSVB/Homo_sapiens/USA/LA2_82/2013			+	+	+		+		+	+	+	+	+					

strain_name	81	86	103	135	209	228	237	250	251	294	308	318	clade	subclade
RSVA/Homo_sapiens/USA/LA2_18/2013			NLSG	NTTT			NITK	NSTT				NITD	2	2
RSVA/Homo_sapiens/USA/LA2_19/2013			NLSG	NTTT			NITK	NSTT				NITD	2	2
RSVA/Homo_sapiens/USA/LA2_38/2012			NLSG	NTTT			NITK	NSTT				NITD	2	2
RSVA/Homo_sapiens/USA/LA2_67/2013			NLSG	NTTT			NITK	NSTT				NITD	2	2
RSVA/Homo_sapiens/USA/LA2_84/2013			NLSG	NTTT			NITK	NSTT				NITD	2	2
RSVA/Homo_sapiens/USA/LA2_04/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_07/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_08/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_10/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_100/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_103/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_105/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_106/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_11/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_13/2012			NLSG	NTTI			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_15/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_17/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_22/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_28/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_34/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_44/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_49/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_55/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_56/2013			NLSG				NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_62/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_69/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_72/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_73/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_74/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_77/2013			NLSG				NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_83/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_85/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_87/2013			NLSG	NTTT			NTTK					NITK	1	1a
RSVA/Homo_sapiens/USA/LA2_90/2012			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_91/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_94/2013			NLSG	NTTT	NTTK		NTTR						1	1a
RSVA/Homo_sapiens/USA/LA2_95/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_97/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_98/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_99/2013			NLSG	NTTT			NTTK						1	1a
RSVA/Homo_sapiens/USA/LA2_01/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_02/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_05/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_09/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_14/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_21/2013			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_26/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b

strain_name	81	86	103	135	209	228	237	250	251	294	308	318	clade	subclade
RSVA/Homo_sapiens/USA/LA2_29/2013			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_31/2013			NLTG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_33/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_36/2013			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_37/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_40/2013			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_41/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_46/2013			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_48/2013			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_53/2013			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_92/2013			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_27/2012			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_39/2013			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_45/2013			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_78/2013			NLSG	NTTT			NTTK		NTTG			NTTK	1	1b
RSVA/Homo_sapiens/USA/LA2_30/2012			NLSG	NTTT			NTTK		NTTG				1	1c
RSVA/Homo_sapiens/USA/LA2_60/2013			NLSG	NTTT			NTTK		NTTG				1	1c
RSVB/Homo_sapiens/USA/LA2_42/2013		NITT	NQTT			NPTK				NSTQ			3	3a
RSVB/Homo_sapiens/USA/LA2_16/2012											NSTQ		3	3b
RSVB/Homo_sapiens/USA/LA2_24/2012											NSTQ		3	3b
RSVB/Homo_sapiens/USA/LA2_25/2013											NSTQ		3	3b
RSVB/Homo_sapiens/USA/LA2_50/2013	NHTE										NSTQ		3	3b
RSVB/Homo_sapiens/USA/LA2_51/2013											NSTQ		3	3b
RSVB/Homo_sapiens/USA/LA2_82/2013											NSTQ		3	3b

strain_name	g_duplication	constellation_95
RSVA/Homo_sapiens/USA/LA2_18/2013	0	15
RSVA/Homo_sapiens/USA/LA2_19/2013	0	15
RSVA/Homo_sapiens/USA/LA2_38/2012	0	15
RSVA/Homo_sapiens/USA/LA2_67/2013	0	15
RSVA/Homo_sapiens/USA/LA2_84/2013	0	15
RSVA/Homo_sapiens/USA/LA2_04/2012	1	1
RSVA/Homo_sapiens/USA/LA2_07/2012	1	1
RSVA/Homo_sapiens/USA/LA2_08/2012	1	1
RSVA/Homo_sapiens/USA/LA2_10/2012	1	1
RSVA/Homo_sapiens/USA/LA2_100/2013	1	1
RSVA/Homo_sapiens/USA/LA2_103/2012	1	1
RSVA/Homo_sapiens/USA/LA2_105/2012	1	1
RSVA/Homo_sapiens/USA/LA2_106/2012	1	1
RSVA/Homo_sapiens/USA/LA2_11/2012	1	1
RSVA/Homo_sapiens/USA/LA2_13/2012	1	1
RSVA/Homo_sapiens/USA/LA2_15/2012	1	1
RSVA/Homo_sapiens/USA/LA2_17/2013	1	1
RSVA/Homo_sapiens/USA/LA2_22/2012	1	1
RSVA/Homo_sapiens/USA/LA2_28/2012	1	1
RSVA/Homo_sapiens/USA/LA2_34/2012	1	1
RSVA/Homo_sapiens/USA/LA2_44/2013	1	1
RSVA/Homo_sapiens/USA/LA2_49/2013	1	1
RSVA/Homo_sapiens/USA/LA2_55/2013	1	1
RSVA/Homo_sapiens/USA/LA2_56/2013	1	1
RSVA/Homo_sapiens/USA/LA2_62/2013	1	1
RSVA/Homo_sapiens/USA/LA2_69/2013	1	1
RSVA/Homo_sapiens/USA/LA2_72/2013	1	1
RSVA/Homo_sapiens/USA/LA2_73/2013	1	1
RSVA/Homo_sapiens/USA/LA2_74/2013	1	1
RSVA/Homo_sapiens/USA/LA2_77/2013	1	1
RSVA/Homo_sapiens/USA/LA2_83/2013	1	1
RSVA/Homo_sapiens/USA/LA2_85/2013	1	1
RSVA/Homo_sapiens/USA/LA2_87/2013	1	1
RSVA/Homo_sapiens/USA/LA2_90/2012	1	1
RSVA/Homo_sapiens/USA/LA2_91/2013	1	1
RSVA/Homo_sapiens/USA/LA2_94/2013	1	1
RSVA/Homo_sapiens/USA/LA2_95/2013	1	1
RSVA/Homo_sapiens/USA/LA2_97/2013	1	1
RSVA/Homo_sapiens/USA/LA2_98/2013	1	1
RSVA/Homo_sapiens/USA/LA2_99/2013	1	1
RSVA/Homo_sapiens/USA/LA2_01/2012	0	3
RSVA/Homo_sapiens/USA/LA2_02/2012	0	3
RSVA/Homo_sapiens/USA/LA2_05/2012	0	3
RSVA/Homo_sapiens/USA/LA2_09/2012	0	3
RSVA/Homo_sapiens/USA/LA2_14/2012	0	3
RSVA/Homo_sapiens/USA/LA2_21/2013	0	3
RSVA/Homo_sapiens/USA/LA2_26/2012	0	3

strain_name	g_duplication	constellation_95
RSVA/Homo_sapiens/USA/LA2_29/2013	0	3
RSVA/Homo_sapiens/USA/LA2_31/2013	0	3
RSVA/Homo_sapiens/USA/LA2_33/2012	0	3
RSVA/Homo_sapiens/USA/LA2_36/2013	0	3
RSVA/Homo_sapiens/USA/LA2_37/2012	0	3
RSVA/Homo_sapiens/USA/LA2_40/2013	0	3
RSVA/Homo_sapiens/USA/LA2_41/2012	0	3
RSVA/Homo_sapiens/USA/LA2_46/2013	0	3
RSVA/Homo_sapiens/USA/LA2_48/2013	0	3
RSVA/Homo_sapiens/USA/LA2_53/2013	0	3
RSVA/Homo_sapiens/USA/LA2_92/2013	0	3
RSVA/Homo_sapiens/USA/LA2_27/2012	1	1
RSVA/Homo_sapiens/USA/LA2_39/2013	1	1
RSVA/Homo_sapiens/USA/LA2_45/2013	1	1
RSVA/Homo_sapiens/USA/LA2_78/2013	1	1
RSVA/Homo_sapiens/USA/LA2_30/2012	0	3
RSVA/Homo_sapiens/USA/LA2_60/2013	0	3
RSVB/Homo_sapiens/USA/LA2_42/2013	1	22
RSVB/Homo_sapiens/USA/LA2_16/2012	1	19
RSVB/Homo_sapiens/USA/LA2_24/2012	1	19
RSVB/Homo_sapiens/USA/LA2_25/2013	1	19
RSVB/Homo_sapiens/USA/LA2_50/2013	1	19
RSVB/Homo_sapiens/USA/LA2_51/2013	1	19
RSVB/Homo_sapiens/USA/LA2_82/2013	1	19

Table 21 dN/dS Results for Positive or Diversifying Selection Sites Across All RSV Genes Categorized by RSV-A and RSV-B. Orange shading indicates statistically significant results at default thresholds of 0.1.

Gene	Codon	SLAC dN-dS	SLAC p-value	FEL dN-dS	FEL p-value	MEME ω+	MEME p-value	FUBAR dN-dS	FUBAR Post. Pr
F RSV-A	23	10.64	0.059	3.942	0.027	>100	0.072	0.619	0.937
	117	6.071	0.233	4.39	0.04	>100	0.098	0.499	0.874
	518	7.6	0.132	3.038	0.054	>100	0.119	0.317	0.859
	560	-12.241	0.986	-4.579	0.154	>100	0.01	-1.091	0.03
	563	4.151	0.467	2.928	0.345	>100	0.004	0.071	0.606
	574	3.596	0.447	3.979	0.098	2.668	0.326	0.565	0.854
F RSV-B	12	4.279	0.342	4.372	0.063	>100	0.093	0.232	0.79
	15	-0.011	0.742	0.372	0.906	29.882	0.091	-0.064	0.381
	45	6.132	0.183	7.358	0.065	>100	0.088	0.516	0.848
	292	2.277	0.768	3.377	0.429	>100	0.096	-0.013	0.545
	516	2.991	0.445	3.106	0.305	>100	0.017	0.054	0.585
G RSV-A	15	2.258	0.252	1.161	0.052	>100	0.075	0.188	0.855
	71	4.335	0.035	0.947	0.067	>100	0.07	0.172	0.862
	94	3.131	0.229	1.929	0.005	>100	0.019	0.436	0.961
	101	5.909	0.046	1.463	0.053	>100	0.022	0.266	0.905
	117	5.929	0.057	2.115	0.089	>100	0.025	0.569	0.933
	126	5.224	0.048	1.311	0.116	5.904	0.062	0.292	0.874
	132	2.212	0.3	1.401	0.041	>100	0.02	0.237	0.873
	142	-0.156	0.764	-0.276	0.838	>100	0.035	-0.087	0.409
	161	1.252	0.554	0.682	0.087	>100	0.161	0.135	0.81
	196	2.957	0.199	1.961	0.016	>100	0.022	0.399	0.929
	206	2.147	0.303	0.786	0.248	>100	0.068	0.068	0.668
	215	4.196	0.088	1.307	0.054	>100	0.069	0.197	0.868
	231	-0.699	0.889	-0.706	0.412	>100	0.008	-0.162	0.231
	241	4.912	0.057	2.124	0.055	>100	0.049	0.511	0.931
	244	6.76	0.041	3.218	0.046	>100	0.065	0.824	0.976
	250	1.788	0.426	0.629	0.618	>100	0.041	0.236	0.697
	255	1.879	0.413	1.168	0.028	>100	0.069	0.21	0.89
	256	5.836	0.019	1.826	0.031	>100	0.029	0.39	0.931
	258	4.576	0.095	2.101	0.027	>100	0.057	0.45	0.932
	284	27.952	0	-0.274	0.343	0	0.67	-0.094	0.211
	286	5.094	0.127	2.393	0.043	14.108	0.044	0.763	0.983
	293	-2.099	0.855	-1.181	0.335	>100	0.019	-0.401	0.155
	298	7.731	0.014	2.424	0.088	24.998	0.053	0.758	0.969
	299	-1.376	0.861	-0.565	0.615	>100	0.001	-0.169	0.309

	310	10.06	0	2.192	0.005	>100	0.005	0.701	0.99
	314	7.833	0.006	2.619	0.014	>100	0.011	0.77	0.988
	321	4.579	0.183	2.661	0.142	33.497	0.047	0.83	0.937
	322	8.797	0.206	7.919	0.01	49.085	0.001	3.288	0.997
G RSV-B	121	-0.346	0.801	2.167	0.502	23.176	0.082	0.116	0.594
	133	2.984	0.198	6.951	0.048	>100	0.074	0.398	0.84
	138	-1.492	0.889	-1.283	0.706	>100	0.002	-0.146	0.33
	154	1.393	0.51	3.998	0.076	>100	0.12	0.235	0.791
	159	11.482	0.006	0	1	0	0.67	-0.085	0.343
	207	2.984	0.198	5.871	0.064	>100	0.088	0.501	0.888
	219	3.735	0.08	5.317	0.121	>100	0.137	0.403	0.856
	223	1.348	0.472	4.418	0.245	26.301	0.041	-0.089	0.336
	236	14.023	0	0	1	0	0.67	1.808	0.959
	272	6.114	0.241	13.155	0.101	>100	0.078	0.413	0.847
	275	5.409	0.299	5.827	0.077	>100	0.111	0.429	0.796
	283	2.719	0.324	8.765	0.194	>100	0.029	1.394	0.976
	292	4.343	0.128	10.214	0.026	>100	0.053	0.052	0.592
	302	2.965	0.203	4.593	0.094	>100	0.122	0.385	0.862
	323	8.764	0.345	14.761	0.081	>100	0.116	1.478	0.856
L RSV-A	171	-5.405	0.963	-4.598	0.115	>100	0.035	-0.828	0.06
	218	5.187	0.335	3.83	0.036	>100	0.1	0.313	0.8
	400	-9.989	0.975	-8.375	0.247	>100	0.085	-1.027	0.181
	746	1.425	0.631	1.958	0.395	>100	0	0.067	0.546
	1049	3.458	0.483	2.218	0.104	>100	0.034	0.154	0.679
	1161	3.109	0.442	3.659	0.289	>100	0.008	-0.014	0.468
	1168	3.603	0.444	1.798	0.281	>100	0.012	0.023	0.535
	1313	-3.308	0.938	-5.002	0.253	>100	0.049	-0.686	0.126
	1490	-10.399	0.992	-9.509	0.056	>100	0.033	-1.847	0.021
	1592	3.603	0.47	2.908	0.074	>100	0	0.207	0.716
	1724	3.468	0.547	2.956	0.356	>100	0.004	0.008	0.539
	1725	13.929	0.056	8.718	0.035	>100	0.105	1.499	0.957
	1874	-0.871	0.8	-3.307	0.49	>100	0.08	-0.508	0.24
	1948	-6.352	0.974	-5.264	0.133	>100	0.089	-0.962	0.058
	1967	-1.802	0.889	-1.605	0.508	>100	0.088	-0.385	0.165
	2039	3.595	0.446	2.524	0.082	>100	0.136	0.19	0.712
L RSV-B	102	7.131	0.41	35.665	0.256	>100	0.001	1.115	0.718
	376	4.577	0.483	18.206	0.08	>100	0.182	0.411	0.72
	1733	85.758	0	-20.487	0.706	0	0.67	-0.76	0.053
M RSV-A	NONE								
M RSV-B	NONE								
M2-1	117	4.042	0.241	7.118	0.026	>100	0.12	0.531	0.919

RSV-A	170	2	0.501	2.718	0.1	>100	0.258	0.114	0.747
M2-1 RSV-B	120	3.844	0.444	9.803	0.147	>100	0.092	0.117	0.723
	142	3.729	0.472	22.741	0.055	>100	0.223	0.498	0.826
M2-2 RSV-A	69	4.196	0.137	6.029	0.067	>100	0.08	0.411	0.92
M2-2 RSV-B	52	8.421	0.255	25.323	0.211	>100	0.214	0.629	0.905
N RSV-A	216	7.614	0.107	6.968	0.018	>100	0.069	0.804	0.966
N RSV-B	NONE								
NS1 RSV-A	NONE								
NS1 RSV-B	NONE								
NS2 RSV-A	6	2.459	0.274	6.483	0.047	>100	0.07	0.442	0.9
NS2 RSV-B	NONE								
P RSV-A	73	4.562	0.027	6.157	0.107	>100	0.082	0.713	0.965
P RSV-B	NONE								
SH RSV-A	49	8.459	0.136	4.427	0.18	>100	0.091	0.196	0.848
SH RSV-B	49	9.116	0.132	26.073	0.167	>100	0.144	1.061	0.955

Bibliography

1. Stockwell TB. 2015.
2. Racaniello V. Viruses. iTunes U; 2013.
3. Nederbragt L. Development in High Throughput Sequencing. In: Developments_in_next_generation_sequencing.jpg, editor. Wikipedia2014.
4. Krauss S, Stucker KM, Schobel SA. Long-term surveillance of H7 influenza viruses in American wild aquatic birds: are the H7N3 influenza viruses in wild birds the precursors of highly pathogenic strains in domestic poultry. Emerging microbes & infections. 2015;4:e35.
5. Stucker KM, Schobel SA, Olsen RJ, Hodges HL, Lin X, Halpin RA, et al. Haemagglutinin mutations and glycosylation changes shaped the 2012/13 influenza A(H3N2) epidemic, Houston, Texas. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2015;20(18).
6. Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature. 1976;260(5551):500-7.
7. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage phi X174 DNA. Nature. 1977;265(5596):687-95.
8. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(12):5463-7.
9. Das SR, Hensley SE, Ince WL, Brooke CB, Subba A, Delboy MG, et al. Defining influenza A virus hemagglutinin antigenic drift by sequential monoclonal antibody selection. Cell host & microbe. 2013;13(3):314-23.
10. Djikeng A, Halpin R, Kuzmickas R, Depasse J, Feldblyum J, Sengamalay N, et al. Viral genome sequencing by random priming methods. BMC genomics. 2008;9:5.
11. Stucker KM. 2015.
12. Li K, Shrivastava S, Brownley A, Katzel D, Bera J, Nguyen AT, et al. Automated degenerate PCR primer design for high-throughput sequencing improves efficiency of viral sequencing. Virology journal. 2012;9:261.
13. Zhou B, Donnelly ME, Scholes DT, St George K, Hatta M, Kawaoka Y, et al. Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and Swine origin human influenza a viruses. Journal of virology. 2009;83(19):10309-13.
14. Zhou B, Lin X, Wang W, Halpin RA, Bera J, Stockwell TB, et al. Universal influenza B virus genomic amplification facilitates sequencing, diagnostics, and reverse genetics. Journal of clinical microbiology. 2014;52(5):1330-7.
15. Sommer DD, Delcher AL, Salzberg SL, Pop M. Minimus: a fast, lightweight genome assembler. BMC bioinformatics. 2007;8:64.
16. CLCbio. White paper de novo assembly in CLC Assembly Cell 4.0 2012. Available from: http://www.clcbio.com/files/whitepapers/whitepaper-denovo-assembly-4.pdf.

17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215(3):403-10.
18. CLCbio. White paper on reference assembly in CLC Assembly Cell 3.0 2010. Available from: http://www.clcbio.com/wp-content/uploads/2012/09/white_paper_on_reference_assembly_on_the_CLC_Assembly_Cell.pdf.
19. Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. Nucleic acids research. 1998;26(2):544-8.
20. Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H. Interpolated Markov models for eukaryotic gene finding. Genomics. 1999;59(1):24-31.
21. Wang S, Sundaram JP, Spiro D. VIGOR, an annotation program for small viral genomes. BMC bioinformatics. 2010;11:451.
22. Wang S, Sundaram JP, Stockwell TB. VIGOR extended to annotate genomes for additional 12 different viruses. Nucleic acids research. 2012;40(Web Server issue):W186-92.
23. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, et al. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic acids research. 2012;40(Database issue):D593-8.
24. Hulo C, de Castro E, Masson P, Bougueleret L, Bairoch A, Xenarios I, et al. ViralZone: a knowledge resource to understand virus diversity. Nucleic acids research. 2011;39(Database issue):D576-82.
25. Masson P, Hulo C, De Castro E, Bitter H, Gruenbaum L, Essioux L, et al. ViralZone: recent updates to the virus knowledge resource. Nucleic acids research. 2013;41(Database issue):D579-83.
26. Galperin MY, Rigden DJ, Fernandez-Suarez XM. The 2015 Nucleic Acids Research Database Issue and molecular biology database collection. Nucleic acids research. 2015;43(Database issue):D1-5.
27. Squires RB, Noronha J, Hunt V, Garcia-Sastre A, Macken C, Baumgarth N, et al. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. Influenza and other respiratory viruses. 2012;6(6):404-16.
28. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, et al. The influenza virus resource at the National Center for Biotechnology Information. Journal of virology. 2008;82(2):596-601.
29. Bogner PC, I.; Lipaman, D.; Cox, N. A global initiative on sharing avian flu data. Nature. 2006;442.
30. Brister JR, Bao Y, Kuiken C, Lefkowitz EJ, Le Mercier P, Leplae R, et al. Towards Viral Genome Annotation Standards, Report from the 2010 NCBI Annotation Workshop. Viruses. 2010;2(10):2258-68.
31. Dugan VG, Emrich SJ, Giraldo-Calderon GI, Harb OS, Newman RM, Pickett BE, et al. Standardized metadata for human pathogen/vector genomic sequences. PloS one. 2014;9(6):e99979.
32. Pickett BE, Liu M, Sadat EL, Squires RB, Noronha JM, He S, et al. Metadata-driven comparative analysis tool for sequences (meta-CATS): an automated process for identifying significant sequence variations that correlate with virus attributes. Virology. 2013;447(1-2):45-51.

33. Miller MAP, W.; Schwartz, T. Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees. ACM2010.
34. Gascuel O, Steel M. Neighbor-joining revealed. Molecular biology and evolution. 2006;23(11):1997-2000.
35. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution. 1987;4(4):406-25.
36. Holder M, Lewis PO. Phylogeny estimation: traditional and Bayesian approaches. Nature reviews Genetics. 2003;4(4):275-84.
37. Zwickl DJ. Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets Under the Maximum Likelihood Criterion: The University of Texas at Austin; 2006.
38. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology. 2007;7:214.
39. Li K, Venter E, Yooseph S, Stockwell TB, Eckerle LD, Denison MR, et al. ANDES: Statistical tools for the ANalyses of DEep Sequencing. BMC research notes. 2010;3:199.
40. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Molecular biology and evolution. 2005;22(5):1208-22.
41. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS genetics. 2012;8(7):e1002764.
42. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, et al. Mapping the antigenic and genetic evolution of influenza virus. Science. 2004;305(5682):371-6.
43. Agoti CN, Otieno JR, Munywoki PK, Mwihuri AG, Cane PA, Nokes DJ, et al. Local evolutionary patterns of human respiratory syncytial virus derived from whole-genome sequencing. Journal of virology. 2015;89(7):3444-54.
44. Dusek RJ, Hallgrimsson GT, Ip HS, Jonsson JE, Sreevatsan S, Nashold SW, et al. North Atlantic migratory bird flyways provide routes for intercontinental movement of avian influenza viruses. PloS one. 2014;9(3):e92075.
45. Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS computational biology. 2013;9(3):e1002947.
46. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. Molecular biology and evolution. 2010;27(8):1877-85.
47. Holmes EC, Grenfell BT. Discovering the phylodynamics of RNA viruses. PLoS computational biology. 2009;5(10):e1000505.
48. Dugan VG, Chen R, Spiro DJ, Sengamalay N, Zaborsky J, Ghedin E, et al. The evolutionary genetics and emergence of avian influenza viruses in wild birds. PLoS pathogens. 2008;4(5):e1000076.
49. Fulvini AA, Ramanunninair M, Le J, Pokorny BA, Arroyo JM, Silverman J, et al. Gene constellation of influenza A virus reassortants with high growth phenotype prepared as seed candidates for vaccine production. PloS one. 2011;6(6):e20823.

50. Gonzalez-Reiche AS, Morales-Betoulle ME, Alvarez D, Betoulle JL, Muller ML, Sosa SM, et al. Influenza a viruses from wild birds in Guatemala belong to the North American lineage. PloS one. 2012;7(3):e32873.
51. Lu G, Rowley T, Garten R, Donis RO. FluGenome: a web tool for genotyping influenza A virus. Nucleic acids research. 2007;35(Web Server issue):W275-9.
52. Matthijnssens J, Ciarlet M, Heiman E, Arijs I, Delbeke T, McDonald SM, et al. Full genome-based classification of rotaviruses reveals a common origin between human Wa-Like and porcine rotavirus strains and human DS-1-like and bovine rotavirus strains. Journal of virology. 2008;82(7):3204-19.
53. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics. 2010;26(19):2462-3.
54. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. Automated phylogenetic detection of recombination using a genetic algorithm. Molecular biology and evolution. 2006;23(10):1891-901.
55. Watson SJ, Welkers MR, Depledge DP, Coulter E, Breuer JM, de Jong MD, et al. Viral population analysis and minority-variant detection using short read next-generation sequencing. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2013;368(1614):20120205.
56. Isakov O, Borderia AV, Golan D, Hamenahem A, Celniker G, Yoffe L, et al. Deep sequencing analysis of viral infection and evolution allows rapid and detailed characterization of viral mutant spectrum. Bioinformatics. 2015.
57. Grad YH, Newman R, Zody M, Yang X, Murphy R, Qu J, et al. Within-host whole-genome deep sequencing and diversity analysis of human respiratory syncytial virus infection reveals dynamics of genomic diversity in the absence and presence of immune pressure. Journal of virology. 2014;88(13):7286-93.
58. Parker J, Rambaut A, Pybus OG. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases. 2008;8(3):239-46.
59. Cummings MP, Neel MC, Shaw KL. A genealogical approach to quantifying lineage divergence. Evolution; international journal of organic evolution. 2008;62(9):2411-22.
60. Barr IG, McCauley J, Cox N, Daniels R, Engelhardt OG, Fukuda K, et al. Epidemiological, antigenic and genetic characteristics of seasonal influenza A(H1N1), A(H3N2) and B influenza viruses: basis for the WHO recommendation on the composition of influenza vaccines for use in the 2009-2010 Northern Hemisphere season. Vaccine. 2010;28(5):1156-67.
61. Fukuyama S, Kawaoka Y. The pathogenesis of influenza virus infections: the contributions of virus and host factors. Current opinion in immunology. 2011;23(4):481-6.
62. Bush RM, Fitch WM, Bender CA, Cox NJ. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Molecular biology and evolution. 1999;16(11):1457-65.

63. Smith DJ, Forrest S, Ackley DH, Perelson AS. Variable efficacy of repeated annual influenza vaccination. Proceedings of the National Academy of Sciences of the United States of America. 1999;96(24):14001-6.
64. Rappuoli R, Dormitzer PR. Influenza: options to improve pandemic preparation. Science. 2012;336(6088):1531-3.
65. Shih AC, Hsiao TC, Ho MS, Li WH. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(15):6283-8.
66. Smith DJ. Predictability and preparedness in influenza control. Science. 2006;312(5772):392-4.
67. Huang JW, King CC, Yang JM. Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses. BMC bioinformatics. 2009;10 Suppl 1:S41.
68. Huang JW, Yang JM. Changed epitopes drive the antigenic drift for influenza A (H3N2) viruses. BMC bioinformatics. 2011;12 Suppl 1:S31.
69. Xia Z, Jin G, Zhu J, Zhou R. Using a mutual information-based site transition network to map the genetic evolution of influenza A/H3N2 virus. Bioinformatics. 2009;25(18):2309-17.
70. Lu B, Zhou H, Ye D, Kemble G, Jin H. Improvement of influenza A/Fujian/411/02 (H3N2) virus growth in embryonated chicken eggs by balancing the hemagglutinin and neuraminidase activities, using reverse genetics. Journal of virology. 2005;79(11):6763-71.
71. Pan K, Subieta KC, Deem MW. A novel sequence-based antigenic distance measure for H1N1, with application to vaccine effectiveness and the selection of vaccine strains. Protein engineering, design & selection : PEDS. 2011;24(3):291-9.
72. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic acids research. 2005;33(2):511-8.
73. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. Briefings in bioinformatics. 2008;9(4):286-98.
74. Ward Jr. JH. Hierarchical Grouping to Optimize an Objective Function. JSTOR. 1963;58(301):236-44.
75. Sun S, Wang Q, Zhao F, Chen W, Li Z. Prediction of biological functions on glycosylation site migrations in human influenza H1N1 viruses. PloS one. 2012;7(2):e32119.
76. Kang S, Yang IS, Lee JY, Park Y, Oh HB, Kang C, et al. Epidemiologic study of human influenza virus infection in South Korea from 1999 to 2007: origin and evolution of A/Fujian/411/2002-like strains. Journal of clinical microbiology. 2010;48(6):2177-85.
77. Centers for Disease C, Prevention. Assessment of the effectiveness of the 2003-04 influenza vaccine among children and adults--Colorado, 2003. MMWR Morbidity and mortality weekly report. 2004;53(31):707-10.
78. Ambrose CS, Levin MJ. The rationale for quadrivalent influenza vaccines. Human vaccines & immunotherapeutics. 2012;8(1):81-8.
79. CDC. The 2012-2013 Influenza Season 2013 [7/16/2015]. Available from: http://www.cdc.gov/flu/pastseasons/1213season.htm.

80. CDC. What You Should Know for the 2014-2015 Influenza Season 2015. Available from: http://www.cdc.gov/flu/about/season/flu-season-2014-2015.htm.
81. Skowronski DM, Janjua NZ, De Serres G, Sabaiduc S, Eshaghi A, Dickinson JA, et al. Low 2012-13 influenza vaccine effectiveness associated with mutation in the egg-adapted H3N2 vaccine strain not antigenic drift in circulating viruses. PloS one. 2014;9(3):e92153.
82. Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics. 1992;131(2):479-91.
83. CDC. Trends and Surveillance 2015 [cited 2015 5/7/2015]. Available from: http://www.cdc.gov/rsv/research/us-surveillance.html.
84. WHO. Influenza Fact Sheet. 2015.
85. Hall CB, Simoes EA, Anderson LJ. Clinical and epidemiologic features of respiratory syncytial virus. Current topics in microbiology and immunology. 2013;372:39-57.
86. Nair H, Nokes DJ, Gessner BD, Dherani M, Madhi SA, Singleton RJ, et al. Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children: a systematic review and meta-analysis. Lancet. 2010;375(9725):1545-55.
87. Library PVR. Respiratory syncytial virus (RSV): path.org; 2015 [7/12/2015]. Available from: http://www.path.org/vaccineresources/rsv.php.
88. CDC. 2014 Ebola Outbreak in West Africa 2015 [7/12/2013]. Available from: http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/index.html.
89. WHO. Ebola Situation Report - 8 July 2015 2015 [updated 7/8/20157/12/2015]. Available from: http://apps.who.int/ebola/current-situation/ebola-situation-report-8-july-2015.
90. Neverov AD, Lezhnina KV, Kondrashov AS, Bazykin GA. Intrasubtype reassortments cause adaptive amino acid replacements in H3N2 influenza genes. PLoS genetics. 2014;10(1):e1004037.
91. Maes P, Matthijnssens J, Rahman M, Van Ranst M. RotaC: a web-based tool for the complete genome classification of group A rotaviruses. BMC microbiology. 2009;9:238.
92. Matthijnssens J, Taraporewala ZF, Yang H, Rao S, Yuan L, Cao D, et al. Simian rotaviruses possess divergent gene constellations that originated from interspecies transmission and reassortment. Journal of virology. 2010;84(4):2013-26.
93. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013;30(4):772-80.
94. Hamming RW. Error detecting and error correcting codes. The Bell Systems Technical Journal. 1950;29(2):147-60.
95. Bazinet AL, Zwickl DJ, Cummings MP. A gateway for phylogenetic analysis powered by grid computing featuring GARLI 2.0. Systematic biology. 2014;63(5):812-8.
96. CLCbio. CLC Genomics Workbench 6.5. 2015.
97. Cox N. Seasonal Influenza and Zoonotic Influenza. VRBPAC2012.

98. Holmes EC, Ghedin E, Miller N, Taylor J, Bao Y, St George K, et al. Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. PLoS biology. 2005;3(9):e300.
99. Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345(6202):1369-72.
100. Hoenen T, Safronetz D, Groseth A, Wollenberg KR, Koita OA, Diarra B, et al. Virology. Mutation rate and genotype variation of Ebola virus from Mali case sequences. Science. 2015;348(6230):117-9.
101. Nagarajan N, Kingsford C. GiRaF: robust, computational identification of influenza reassortments via graph mining. Nucleic acids research. 2011;39(6):e34.
102. Defays D. An efficient algorithm for a complete link method. The Computer Journal. 1977;20(4):334-6.
103. Collins PL, Fearns R, Graham BS. Respiratory syncytial virus: virology, reverse genetics, and pathogenesis of disease. Current topics in microbiology and immunology. 2013;372:3-38.
104. Sullender WM, Mufson MA, Anderson LJ, Wertz GW. Genetic diversity of the attachment protein of subgroup B respiratory syncytial viruses. Journal of virology. 1991;65(10):5425-34.
105. Galiano MC, Luchsinger V, Videla CM, De Souza L, Puch SS, Palomo C, et al. Intragroup antigenic diversity of human respiratory syncytial virus (group A) isolated in Argentina and Chile. Journal of medical virology. 2005;77(2):311-6.
106. Tan L, Coenjaerts FE, Houspie L, Viveen MC, van Bleek GM, Wiertz EJ, et al. The comparative genomics of human respiratory syncytial virus subgroups A and B: genetic variability and molecular evolutionary dynamics. Journal of virology. 2013;87(14):8213-26.
107. Homaira N, Rawlinson W, Snelling TL, Jaffe A. Effectiveness of Palivizumab in Preventing RSV Hospitalization in High Risk Children: A Real-World Perspective. International journal of pediatrics. 2014;2014:571609.
108. Zhao X, Sullender WM. In vivo selection of respiratory syncytial viruses resistant to palivizumab. Journal of virology. 2005;79(7):3962-8.
109. Choudhary ML, Anand SP, Wadhwa BS, Chadha MS. Genetic variability of human respiratory syncytial virus in Pune, Western India. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases. 2013;20:369-77.
110. Eshaghi A, Duvvuri VR, Lai R, Nadarajah JT, Li A, Patel SN, et al. Genetic variability of human respiratory syncytial virus A strains circulating in Ontario: a novel genotype with a 72 nucleotide G gene duplication. PloS one. 2012;7(3):e32807.
111. Tan L, Lemey P, Houspie L, Viveen MC, Jansen NJ, van Loon AM, et al. Genetic variability among complete human respiratory syncytial virus subgroup A genomes: bridging molecular evolutionary dynamics and epidemiology. PloS one. 2012;7(12):e51439.
112. Ren L, Xia Q, Xiao Q, Zhou L, Zang N, Long X, et al. The genetic variability of glycoproteins among respiratory syncytial virus subtype A in China between 2009 and 2013. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases. 2014;27:339-47.

113. Trento A, Galiano M, Videla C, Carballal G, Garcia-Barreno B, Melero JA, et al. Major changes in the G protein of human respiratory syncytial virus isolates introduced by a duplication of 60 nucleotides. The Journal of general virology. 2003;84(Pt 11):3115-20.
114. Trento A, Viegas M, Galiano M, Videla C, Carballal G, Mistchenko AS, et al. Natural history of human respiratory syncytial virus inferred from phylogenetic analysis of the attachment (G) glycoprotein with a 60-nucleotide duplication. Journal of virology. 2006;80(2):975-84.
115. Choudhary ML, Wadhwa BS, Jadhav SM, Chadha MS. Complete Genome Sequences of Two Human Respiratory Syncytial Virus Genotype A Strains from India, RSV-A/NIV1114046/11 and RSV-A/NIV1114073/11. Genome announcements. 2013;1(4).
116. Lee WJ, Kim YJ, Kim DW, Lee HS, Lee HY, Kim K. Complete genome sequence of human respiratory syncytial virus genotype A with a 72-nucleotide duplication in the attachment protein G gene. Journal of virology. 2012;86(24):13810-1.
117. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nature methods. 2012;9(8):772.
118. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular biology and evolution. 2012;29(8):1969-73.
119. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Molecular biology and evolution. 2013;30(3):713-24.
120. Ferreira MAR, Suchard MA. Bayesian analysis of elapsed times in continuous-time Markov chains. Can J Stat. 2008;36(3):355-68.
121. R. Gupta EJaSB. NetNGlyc 1.0 Server 2004 [cited 2015]. Available from: http://www.cbs.dtu.dk/services/NetNGlyc/abstract.php.
122. Hansen JE, Lund O, Tolstrup N, Gooley AA, Williams KL, Brunak S. NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. Glycoconjugate journal. 1998;15(2):115-30.
123. Wickham H. ggplot2: elegant graphics for data analysis: Springer New York; 2009.
124. Team RC. R: A Language and Environment for Statistical Computing. 2013.
125. Liu LL, Gallaher MM, Davis RL, Rutter CM, Lewis TC, Marcuse EK. Use of a respiratory clinical score among different providers. Pediatric pulmonology. 2004;37(3):243-8.
126. Tregoning JS, Schwarze J. Respiratory viral infections in infants: causes, clinical symptoms, virology, and immunology. Clinical microbiology reviews. 2010;23(1):74-98.
127. Johnson PR, Spriggs MK, Olmsted RA, Collins PL. The G glycoprotein of human respiratory syncytial viruses of subgroups A and B: extensive sequence divergence between antigenically related proteins. Proceedings of the National Academy of Sciences of the United States of America. 1987;84(16):5625-9.
128. Sullender WM. Respiratory syncytial virus genetic and antigenic diversity. Clinical microbiology reviews. 2000;13(1):1-15, table of contents.

129. Hotard AL, Laikhter E, Brooks K, Hartert TV, Moore ML. Functional Analysis of the 60 Nucleotide Duplication in the Respiratory Syncytial Virus Buenos Aires Strain Attachment Glycoprotein. Journal of virology. 2015.
130. Forcic D, Ivancic-Jelecki J, Mlinaric-Galinovic G, Vojnovic G, Babic-Erceg A, Tabain I. A study of the genetic variability of human respiratory syncytial virus in Croatia, 2006-2008. Journal of medical virology. 2012;84(12):1985-92.
131. McLellan JS, Ray WC, Peeples ME. Structure and function of respiratory syncytial virus surface glycoproteins. Current topics in microbiology and immunology. 2013;372:83-104.
132. Melero JA, Moore ML. Influence of respiratory syncytial virus strain differences on pathogenesis and immunity. Current topics in microbiology and immunology. 2013;372:59-82.
133. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology. 2007;25(11):1251-5.
134. Krishnakumar V, Hanlon MR, Contrino S, Ferlanti ES, Karamycheva S, Kim M, et al. Araport: the Arabidopsis information portal. Nucleic acids research. 2015;43(Database issue):D1003-9.
135. Miller MA, Schwartz T, Pickett BE, He S, Klem EB, Scheuermann RH, et al. A RESTful API for Access to Phylogenetic Tools via the CIPRES Science Gateway. Evolutionary bioinformatics online. 2015;11:43-8.
136. Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, et al. Data integration in the era of omics: current and future challenges. BMC systems biology. 2014;8 Suppl 2:I1.
137. Hay SI, George DB, Moyes CL, Brownstein JS. Big data opportunities for global infectious disease surveillance. PLoS medicine. 2013;10(4):e1001413.
138. Kachigan SK. Multivariate Statistical Analysis: A Conceptual Introduction, 2nd Edition: Radius Press; 1991.
139. Costello JC, Heiser LM, Georgii E, Gonen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nature biotechnology. 2014;32(12):1202-12.
140. De Angelis D, Presanis AM, Birrell PJ, Tomba GS, House T. Four key challenges in infectious disease modelling using data from multiple sources. Epidemics. 2015;10:83-7.