# DEFENSE THREAT REDUCTION AGENCY
## Scientific & Technical Review Information

2011 020

PA CONTROL NUMBER: 21 June 11 PA-11-233

PM / PHONE / EMAIL: Nicole Whealen 703.767.6354

BRANCH CHIEF / PHONE /EMAIL:

DIVISION CHIEF / PHONE:

DIRECTORATE / DIRECTOR /PHONE: David Hamon 703.767.5713

ENTERPRISE / OFFICE / PHONE:

PUBLIC AFFAIRS: Richard M. Cole (Chief, PA)

SUSPENSE: 17 Jul 2011

DATE: 17 Jun 2011

DATE:

DATE:

DATE:

DATE:

DATE: 6/30/11

1. TTLE: Anticipating Viral Species Jumps: Bioinformatics and Data Needs CONTRACT NUMBER

ORIGINATOR: Flanagan, Meg; Leighton, Terrance; Dudley, Joseph

2. TYPE OF MATERIAL: ☒ PAPER ☐ PRESENTATION ☐ ABSTRACT ☐ OTHER

3. OVERALL CLASSIFICATION: ☒ CONTRACTOR UNCLASS ☒ PROJECT MANAGER UNCLASS

A. Review authority for unclassified material is the responsibility of the PM. Your signature indicates the material has undergone technical and security review.

B. Warning Notices/Caveats: ☐ RD ☐ FRD ☐ CNWDI ☐ NATO RELEASABLE ☐ SUBJECT TO EXPORT CONTROL LAWS

C. Distribution Statement:

☒ A. Approved for public release; distribution is unlimited (unclassified papers only).

CLEARED
for public release

JUN 30 2011

PA Opns
Defense Threat Reduction Agency

☐ B. Distribution authorized to U.S. Government agencies only; (check the following):

- ☐ Contractor Performance Evaluation
- ☐ Foreign Government Information
- ☐ Administrative or Operational Use
- ☐ Specific Authority
- ☐ Premature Dissemination
- ☐ Proprietary Information
- ☐ Test and Evaluation
- ☐ Software Documentation
- ☐ Critical Technology

☐ C. Distribution authorized to U.S. Government agencies and their contractors: (check the following):

- ☐ Critical Technology
- ☐ Specific Authority
- ☐ Administrative or Operational Use
- ☐ Software Documentation
- ☐ Foreign Government Information

☐ D. Distribution authorized to the Department of Defense and U.S. DoD Contractors only; (check the following):

- ☐ Foreign Government Information
- ☐ Critical Technology
- ☐ Administrative or Operational Use
- ☐ Software Documentation
- ☐ Foreign Government Information

☐ E. Distribution authorized to DoD Components only; (check the following):

- ☐ Administrative or Operational Use
- ☐ Premature Dissemination
- ☐ Critical Technology
- ☐ Foreign Government Information
- ☐ Direct Military Support
- ☐ Software Documentation
- ☐ Specific Authority
- ☐ Proprietary Information
- ☐ Test and Evaluation
- ☐ Contractor Performance Evaluation

☐ F. Further dissemination only as directed.

☐ X. Distribution authorized to U.S. Government agencies and private individuals or enterprises eligible to obtain export-controlled technical data in accordance with DoD Directive 5230.25 (unclassified papers only).

APPROVED JTW 23 June 2011
OPSEC REVIEW/DATE

4. MATERIAL TO BE: ☐ Presented ☐ Published Date Required:

Name of Conference or Journal:

Remarks:

# Anticipating Viral Species Jumps:
## Bioinformatics and Data Needs

Written by:
Meg Flanagan
*The Pennsylvania State University*

Terrance Leighton
*SAIC*

and

Joseph Dudley
*SAIC*

**June 2011**

---

This report is the product of collaboration between the Defense Threat Reduction Agency's Office of Strategic Research and Dialogues and Science Applications International Corporation, Inc.

The views expressed herein are those of the authors and do not necessarily reflect the official policy or position of the Defense Threat Reduction Agency, the Department of Defense, or the United States Government.




**Defense Threat Reduction Agency**
**Office of Strategic Research and Dialogues**
Report Number OSRD 2011 020
Contract/MIPR Number 01-03-D-0017
Project Cost: $100,000

The mission of the Defense Threat Reduction Agency (DTRA) is to safeguard America and its allies from weapons of mass destruction (chemical, biological, radiological, nuclear, and high explosives) by providing capabilities to reduce, eliminate, and counter the threat, and mitigate its effects.

The Office of Strategic Research and Dialogues (OSRD) supports this mission by providing long-term rolling horizon perspectives to help DTRA leadership identify, plan, and persuasively communicate what is needed in the near term to achieve the longer-term goals inherent in the agency's mission. OSRD also emphasizes the identification, integration, and further development of leading strategic thinking and analysis on the most intractable problems related to combating weapons of mass destruction.

For further information on this project, or on OSRD's broader research program, please contact:

Defense Threat Reduction Agency
Office of Strategic Research and Dialogues
8725 John J. Kingman Road
Ft. Belvoir, VA 22060-6201

OSRDInfo@dtra.mil

# Anticipating Viral Species Jumps: Bioinformatics & Data Needs

Meg L. Flanagan, Ph.D., The Pennsylvania State University
Terrance J. Leighton, Ph.D., Science Applications International Corporation
Joseph P. Dudley, Ph.D., Science Applications International Corporation

DTRA/SCC-WMD

## EXECUTIVE SUMMARY

Viral species jumps (also called host jumps) occur when a virus acquires the ability to infect and spread among individuals of a new host species. Historical examples of animal viruses that jumped into human hosts include HIV, SARS coronavirus and influenza A virus. Globally, these viruses have exacted high socioeconomic and health costs. The ability to predict viral species jumps can reduce such costs by enabling swifter outbreak mitigation strategies and prevention of initial or secondary human infection. Currently, most emerging infectious disease surveillance efforts seek the *ecological drivers behind spillover events* – factors like climate, land use and population migrations driving infections that do not spread between humans. By contrast, we focus here on the *evolutionary drivers behind species jumps* – the genetic changes over time driving infections that spread efficiently among humans. We see an opportunity to apply field surveillance and laboratory data to better understand how viral species jumps occur. There are publicly available extant data that can be marshaled. To build a mechanistic framework of understanding, data must be integrated and accessible to users for analysis and modeling, as well as formulation and testing of hypotheses. In short, bioinformatics must be applied. To that end, the Defense Threat Reduction Agency's Advanced Systems and Concepts Office hosted a workshop that gathered computational biologists and information scientists to explore the types of data needed, the computational methods required, and suitable platforms to share information among interdisciplinary stakeholders. Three key recommendations arose from the expert presentations given at the workshop and the discussions they prompted:

**Recommendation 1**: **Develop a federated data repository and computational workbench containing extant field and laboratory datasets linked to a toolbox for analyzing viral species jump drivers and mechanisms.** This repository-workbench platform, referred to herein as SJOne, should fulfill three core functions: 1) *collation* of disparate data types to enable deeper understanding (and eventual prediction) of species jumps; 2) provision of tools that support multi-level *data analysis and modeling* across disparate stakeholder groups; and 3) development of tools to *visualize* data analyses and *incentivize* their sharing. Since some of the input data already exist or are openly available, SJOne will leverage prior S&T investments and no-cost data streams. Government, academic and private sector users will be able to bridge programmatic silos, increase species jump situational awareness, aggregate data required for analysis and modeling in one location, and collaborate among communities that do not normally interact.

**Recommendation 2**: **Devise statistically supported field sampling plans to detect evolutionary drivers that could precede viral species jumps.** These plans can be deployed *proactively* (at viral "hotspots" where spillover has occurred and jumps are anticipated) and *reactively* (for rapid deployment at early outbreak stages where viral species jumps are nascent or suspected). While the primary objective of a traditional outbreak response is to prevent loss of life, the primary objective of the proposed field sampling plans would be to acquire data that are only available from animal and human hosts near the time of a species jump. As such, these field sampling plans should be deployed independently from traditional responses yet share relevant data with traditional responders.

**Recommendation 3**: **The community should seek funding for pilot studies that test hypotheses of how viral species jumps occur.** Ripe for study are the genetic and phenotypic differences between *Nipah and Nipah-like viruses* in pig, bat and human hosts as well as differences between *Ebola virus* strains that are pathogenic or non-pathogenic in humans. The community should also seek funding for pilot studies to test the hypotheses that smaller viral genomes pose increased risk for species jumps, and that evidence of *horizontal gene transfer* increases species jump risk.

## INTRODUCTION

Viral species jumps (also called host jumps) occur when a virus acquires the ability to infect and spread among individuals of a new host species. Historical examples of animal viruses that jumped into human hosts include HIV (from chimpanzees), SARS coronavirus (from bats) and influenza A virus (from birds). Globally, these viruses have exacted high socioeconomic and health costs. The ability to predict viral species jumps may reduce such costs by enabling swifter outbreak mitigation strategies and preventing initial or secondary human infections. In the long term, developing a deeper understanding of how species jumps occur may enable detection of animal viruses that have a higher risk of jumping to human hosts. In this case, human infections could be reduced or prevented by culling infected animals, deploying prophylactic drugs or restricting contacts between infected animals and humans.

The idea that viral species jumps can be predicted is a contentious one, and several of the underlying disagreements are semantic. First is the distinction between species jumps and spillover events. A **species jump** is defined by *efficient spread by a pathogen within a new host species that was not previously susceptible to that pathogen*. Viruses attain the ability to jump species through the acquisition of genetic changes that allow infection and successful replication within new host cells, as well as successful transmission to other hosts or vector organisms (Parrish *et al*. 2008). By contrast, a **spillover event** is defined by *episodic infection within a non-typical host species by a pathogen that cannot spread sustainably among individuals of that non-typical host*. Thus, sustained spread (or transmission) does not occur in spillover hosts, and outbreaks will not occur if pathogen transmission from the animal reservoir is prevented (Palmer 2007, Nugent 2011). Species jumps from animals to humans therefore pose greater risks than spillover of **zoonotic** pathogens (animal pathogens that are infectious to humans) because spillover events are self-limiting, while species-jumping pathogens can spread from human to human. Despite the higher risk they pose, species jumps are far rarer than spillover events, although spillover events may subsequently evolve into species jumps.

Because spillover events are more common, most current emerging infectious disease (EID) surveillance programs have been designed to detect spillover events into human populations, and even to predict them by detecting zoonotic pathogens in animal reservoirs. The Armed Forces Health Surveillance Center's Global Emerging Infections Surveillance and Response System (AFHSC-GEIS) and the United

**Species jump**

Efficient spread by a pathogen within a new host species that was not previously susceptible to that pathogen (Parrish *et al.* 2008).

**Spillover event**

Episodic infection within a non-typical host species by a pathogen that cannot spread sustainably among individuals of that non-typical host (Nugent 2011).

**Zoonotic**

Adj. for zoonosis: Any disease or infection that is naturally transmissible from vertebrate animals to humans and vice-versa (Pan American Health Organization definition).

**Ecological driver**

An ecological element that causes a change in an organism, community, ecosystem, or other ecological component of the landscape (EPA definition).

**Evolutionary driver**

A genetically determined structure with a function of propelling or steering the evolution of a gene, phenotypic trait or species (Prakash 2008).

**Bioinformatics**

Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data (NIH definition).

States Agency for International Development (USAID) have funded major initiatives to improve global detection and identification of emerging pathogens in humans and animals, and to enhance disease outbreak response capabilities (IOM 2007; IOM 2009; Sueker *et al*. 2010; Russell *et al*. 2011). Much of this research is concerned with discovering and tracking the **ecological drivers** behind spillover events – factors such as climate, land use and population migrations. In addition, non-governmental organizations funded by USAID and Google.org, such as EcoHealth Alliance and Global Viral Forecasting, Inc., have established international disease surveillance systems in wild animals to monitor the potential for spillover outbreaks.

Further, the term "predict" is itself problematic, because societies have pressing needs for accurate predictions that go unmet (e.g., of earthquakes and volcanic eruptions). A tangible example in daily life is the prediction of weather, made possible by vast amounts of continuously collected time-series data. The types of precipitation are well characterized, and the impact of annual temperature variations is understood. In other words, a framework of understanding exists for weather, and this framework – combined with data collection – enable construction of sophisticated models that give high-resolution, by-the-hour predictions of weather with acceptable margins of error.

*Distinguishing **spillover events** from **species jumps**: Most zoonotic disease surveillance programs detect spillover events – incidence of human infection by known pathogens. Surveillance for species jumps goes beyond pathogen detection – it seeks to detect over time evolutionary changes that indicate viral adaptation to new human hosts. Such changes pose increased risk to humans, especially when they enable human-to-human transmission.*

There are also ongoing efforts funded by the National Institutes of Health Fogarty International Center and the Department of Homeland Security Research and Policy for Infectious Disease Dynamics (RAPIDD) to model the **evolutionary drivers** behind species jumps – the genetic changes that viruses undergo as they acquire the ability to infect new host species. However, no single funding agency has a mandate to develop a surveillance capability to predict viral species jumps – that is, to support the longitudinal field and laboratory research necessary to develop and test hypotheses about how viruses *evolve* to infect and sustainably *spread* among new hosts. Fortunately, field and laboratory data generated by infectious disease ecology studies can help to elucidate how viral species jumps occur. Thus, current progress toward predicting spillover events enables future progress toward predicting species jumps.

To extend the weather analogy further, predictive virus surveillance is at a Farmer's Almanac stage – observing seasonal ecological phenomena (e.g., rainfall, vegetation cover) that drive spillover events and looking retrospectively for evolutionary drivers among viruses that have already jumped to humans. These efforts are necessary but not sufficient to build a framework of understanding for how species jumps occur – more continuous data are needed if prediction of viral species jumps is the goal. Many viral data sets already exist that could illuminate *tropism* (which virus infects which tissues in which species), *virulence* (severity of disease) and *transmissibility* (contagiousness, or potential to spread). With sufficient quantities of relevant data and a mechanistic framework of understanding, more sophisticated models could be built to make testable predictions. In the short term, such models would have significant impact on detecting spillover events, and in the long term they could aid prediction of species jumps.

We see an opportunity to use field surveillance and laboratory data to better understand how viral species jumps occur. There are publicly available extant data that can be marshaled. To build a mechanistic framework of understanding, data must be integrated and accessible to users for analysis and modeling, as well as formulation and testing of hypotheses. In short, **bioinformatics** must be applied. To that end, the Defense Threat Reduction Agency's Advanced Systems and Concepts Office hosted a workshop that gathered computational biologists and information scientists to explore the types of data needed, the computational methods required, and suitable platforms to share information among interdisciplinary stakeholders. Three key recommendations arose from the expert presentations given at the workshop and the discussions they prompted: 1) develop a federated data repository and computational workbench containing extant field and laboratory datasets linked to a toolbox for analyzing viral species jump drivers and mechanisms; 2) devise statistically supported field sampling plans to detect evolutionary drivers that could precede viral species jumps; and 3) fund pilot studies that can fill data gaps in the federated data repository as well as test the validity and predictions obtained from field sampling plans.

## SJONE: DATA REPOSITORY AND WORKBENCH DEVELOPMENT

Across different virus-host systems, there are gaps in basic knowledge (e.g., host receptor identity, tissue tropism, genome maps, host range). There is also uncertainty as to how extant data can best be utilized to guide field and laboratory research as well as make predictions. Traditionally, scientists have communicated via publication in peer-reviewed, subfield-specific journals, with little need or impetus to share information outside their subfields. Developing the capability to predict species jumps will require higher-level data integration and analysis that supports actionable decision making. One way to achieve data integration is to create an informatics framework with defined inputs, outputs, objectives and formatting.

Workshop participant Juliet Pulliam, a RAPIDD Fellow at Fogarty International Center, discussed data needs for predicting viral host jumps. She recommended a database that would integrate field and laboratory data, including registries of virus detections, receptor usage, and virus-host interactions. Figure 1 lists examples of these types of data, represented as inputs (purple boxes). These inputs would be integrated within a **data repository**, and disparate stakeholders (green boxes) could retrieve the information they need in the format they want via a **workbench** user interface. Many of the inputs are fragmented or incomplete (red text). Therefore, curating these inputs would reveal data gaps that stakeholders could identify and use to inform their funding decisions. In the short term, this data repository could be circumscribed to prioritized virus-host datasets to provide proof of principle and to test algorithms and metadata requirements. For example, workshop participant John Yin, Professor at the University of Wisconsin-Madison, has built a model of vesicular stomatitis virus (VSV) infection kinetics using parameters from the scientific literature (Lim *et al.* 2006). (VSV is a pathogen of major agricultural importance because it causes symptoms that are clinically indistinguishable from those caused by foot-and-mouth disease virus in hoofed animals.) In the long term, the data repository could be expanded to include data from additional virus, host, vector and reservoir species.

### *Data Acquisition*

Text Mining the Scientific Literature. Obtaining input data (Figure 1) will require a three-pronged strategy. First, the scientific literature contains decades' worth of research on viruses, animal hosts, vectors, and their interactions in the environment

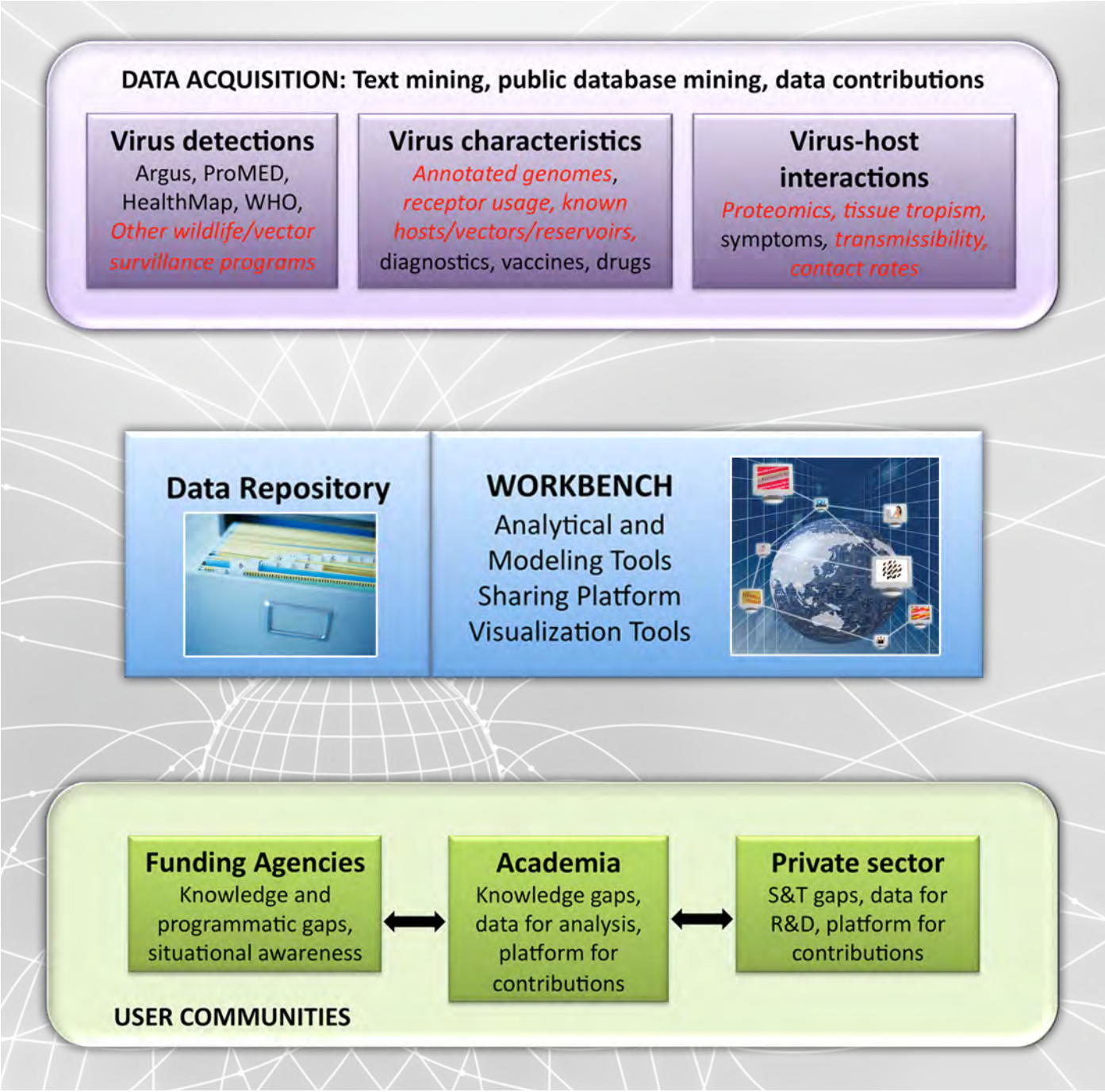



**Figure 1. Notional SJOne informatics framework to understand and make predictions about species jumps.** **Purple boxes** represent types of input data that can be acquired by text mining, database mining, and data contributions from users. Inputs shown in **red text** may be incomplete for a given virus-host system. **Blue boxes** represent the SJOne informatics platform comprised of a data repository and a workbench. The curated data repository stores data acquired from the scientific literature, public databases and contributions. The workbench is the user interface that provides tools to analyze, share and visualize data; to build and test models; and to upload new data into the user workspace. **Green boxes** represent user communities and the types of benefits they can extract from SJOne.

(ecology). Within these research reports are data that are relevant to viral species jumps – for example, the protein receptors that viruses use to enter host cells, or sequences from viral evolution experiments. However, the relevant data are stovepiped among specialty journals that cater to distinct disciplines (e.g., ecologists, virologists, evolutionary biologists) with limited cross-disciplinary interactions. Information isolated within sub-disciplinary journals hampers progress toward a unified mechanistic framework of understanding for viral species jumps, which will require cross-cutting data. Identifying the relevant extant data will require mining the scientific literature using tools borrowed from information extraction (IE), information retrieval (IR) and text mining fields. For example, PubGene[1] is an IE tool that enables users to search across millions of literature records in the National Library of Medicine's PubMed database for information on genes and their associated proteins (without prior knowledge of synonymous names). An alternate approach is an IR system like Textpresso,[2] which allows researchers to search for keywords that appear within a body (corpus) of sentences extracted from the scientific literature (Muller *et al*. 2004). This exposes researchers to associations they might otherwise miss because it bridges bioscience sub-disciplines in ways that manual review does not permit. Also promising is prospective use of text mining tools like Arrowsmith,[3] which enables users to discover keyword connections between two PubMed searches that they designate. The ability to explore non-obvious connections between disparate sub-disciplines enables investigators to explore new hypotheses, which is a key requirement for discovering how species jumps occur.

Data Mining. Second, there are important data that reside in public databases – for example, nucleic acid sequences in GenBank and the European Molecular Biology Laboratory (EMBL) database. (For an annotated list of genomic databases and analytical resources, see Appendix 1.) Users could retrieve and analyze such data without having to leave SJOne – thus SJOne must acquire public data that users need, either by retrieving those data on-demand from public databases or collecting them in the data repository.

Data Contributions. Third, and most challenging, is solicitation of data contributions by individual field and laboratory teams. A lag exists between the time data are collected and the time when results are published, and even the most expeditious journals require time for peer review. Furthermore, not all collected data appear in published form, so the literature is not representative of all data collected. Therefore, acquisition of raw data donated by the collecting investigators would enable SJOne users to analyze those data in novel ways and build more accurate models informed by extended datasets. While scientific culture does not currently favor raw data sharing, there are indications that the culture will shift in response to 1) new funding requirements (e.g., the National Science Foundation's Data Sharing Policy[4]); 2) new ways to tag donated data to compel recognition for the donor; and 3) changing academic attitudes that place greater value on data sharing when assessing tenure or promotion eligibility. (For pertinent discussion of both data sharing issues and the utility of federated data archives, see Reichman *et al*. 2011.)

1 http://www.pubgene.org/
2 http://www.textpresso.org/
3 http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/start.cgi
4 http://www.nsf.gov/bfa/dias/policy/dmp.jsp

### *Analytical and Modeling Tools*

Once relevant data are collated, they need to be analyzed, and results of these analyses must be visualized (and ideally, shared). We propose development of a computational workbench within SJOne to enable users across disparate stakeholder groups (e.g., academic, government, and private sector scientists) to analyze data and to visualize and share results. Workshop participant Richard Scheuermann, Professor and Director of Biomedical Informatics at the University of Texas Southwestern Medical Center, is Principal Investigator of the Virus Pathogen Resource (ViPR)[5], a Bioinformatics Resource Center sponsored by the National Institute of Allergy and Infectious Diseases (NIAID). The two year-old open access ViPR platform hosts data from 13 virus families, including complete genome sequences and 3-D protein structures. Users can upload, analyze, save and share data using ViPR's workbench. ViPR is modeled on a predecessor platform (the Influenza Research Data repository, IRD[6]) that additionally contains surveillance data from samples collected in the field. In IRD, field sample data can be visualized spatially on a searchable map. Other types of data that exist in IRD and can be applied to ViPR include laboratory-generated data (from cell culture and animal studies) and clinical data. Highly relevant to understanding species jumps is the capability in IRD to examine known functional regions of viral genes and proteins and, using these sequence features (SF), to make comparisons across different viral strains to determine how greatly these SF vary. Scheuermann and colleagues are currently adding SF for dengue, hepatitis C and pox viruses by manually searching the literature; in the future they will begin an automated information extraction process to capture SF from other virus families residing in external data repositories. ViPR is a community resource that could be expanded like its IRD predecessor to include extensive laboratory-generated data from cell culture and animal studies.

Text and data mining tools would be used to acquire data for the curated data repository; in turn, the workbench would provide users with these same tools for intra- and extra-repository mining. In this way, users could mine curated data within the repository, as well as outside data sources, and import these data into their personal workspaces. The strength of this approach lies in the ability to quantify whether such correlations are statistically significant (and hence worthy of further inquiry (Jensen *et al*. 2006). Workshop

participant Hsinchun Chen, Professor and Director of the University of Arizona's Artificial Intelligence Lab, is the developer of the BioPortal information system. BioPortal has been applied to international foot-and-mouth disease (FMD) surveillance using field data sets collected from veterinary reference laboratories and the World Organization for Animal Health (OIE) and curated by members of the University of California Davis FMD Lab. The

5 http://www.viprbrc.org/
6 http://www.fludb.org

FMD BioPortal[7] integrates FMD virus genomic data with these field datasets and uses modeling tools to forecast geographical spread of FMD virus. Also integrated is a News component, which searches web-based news sources (e.g., Google, Yahoo) for FMD outbreak events. While specific to one virus-host system, the FMD BioPortal is an established informatics platform with components that could be applied to SJOne for understanding species jumps.

## FIELD SAMPLING PLAN

There are historical examples of viral species jumps that have profoundly affected human society, the most recent being the emergence of SARS coronavirus (SARS-CoV). Learning from these outbreaks post facto is stymied by the absence of data from the primary events that catalyzed the outbreak. Workshop participant James Lloyd-Smith and colleagues (Pepin *et al*. 2010) pointed out that the evolutionary driver of the SARS-CoV jump cannot be discerned by available data from palm civet and human samples. In order to learn more today about the 2003 SARS-CoV jump, sustained sampling from animal and recipient hosts would have had to be done in the past, contemporaneous with the time of the outbreak. In other words, the opportunity is lost. A more recent paper makes the same point for 2009 pandemic H1N1 influenza A virus – that a lack of surveillance data hampers reconstruction of that virus's origins (Vijaykrishna *et al*. 2011).

There is an unmet need to have well-tested sampling plans that can be executed when human outbreaks first originate from an animal host, as in the case of SARS. The sampling strategy should be capable of identifying and discriminating positive selection (i.e., increases in viral fitness) from adaptive fine-tuning (changes in an already well-adapted virus).[8] Such data would make rigorous modeling of species jumps possible and further the understanding required to make predictions possible.

**Box 1: Species Jumps by SARS Coronavirus**

The process of genetic adaptation for human infection by species-jumping viruses may entail an iterative series of adaptations to different hosts. SARS coronavirus, for example, originated in bats but contains gene sequences derived from both mammalian- and avian-adapted ancestral viruses (Wang & Eaton 2007). The ancestral SARS-like bat virus subsequently jumped to humans after infecting at least two wild mammalian species (palm civet, raccoon dog) that were commercially raised and sold for food in China as well as acquiring a 29-base pair deletion that increased the ability of the SARS virus to infect and transmit in human hosts (Guan *et al*. 2003).

We suggest development of statistically supported field sampling plans to detect evolutionary drivers that could precede viral species jumps. These plans can be deployed proactively (at viral "hotspots" where spillover has occurred and jumps are anticipated) and reactively (for rapid deployment at early outbreak stages where viral species jumps are nascent or suspected). While the primary objective of a traditional outbreak response is to prevent loss of life, the primary objective of the proposed field sampling plans would be to acquire data that are

7 Accessible at http://fmdbioportal.ucdavis.edu/

8 Please see Pepin et al. 2010 for an authoritative discussion of evolutionary drivers of species jumps as well as the data gaps that plague the field.

only available from animal and human hosts near the time of a species jump. As such, these field sampling plans should be deployed independently from traditional responses yet share relevant data with traditional responders.

*Reactive* (post-outbreak) public health sampling is primarily focused on contact tracing and trace-back to an index case. These strategies were designed to identify disease sources, but not to discover the drivers that enabled and preceded outbreaks. Representative sampling of aquatic, terrestrial and captive animal populations is distinct and more challenging than medical contact tracing, in which case the probability of detection ($P_d$) of a pathogen is high since infected patients are readily identified. By contrast, the probability of detection for zoonotic viruses in animal populations is low because the ecobiology (host/reservoir/vector dynamics) of many virus-host systems is not well understood. We suggest that environmental sampling design principles, traditionally focused on discovering chemical and radiological contamination, provide models and strategies that may be suited for environmental sampling of zoonotic viruses. Traditional environmental sampling entails collection of samples that are representative of a large area and that are statistically representative of that area. Representative sampling enables one to draw defensible conclusions about viral abundance and distribution within the larger area (and thus reduce cost).

The advantage of *proactive* (background) sampling done with sufficient sensitivity and over a sustained time period is that a time series of pre-emergent species jump processes, dynamics and signals could be accessed, which would allow *de novo* modeling and simulation of disease emergence. These models could then be tested and refined to evaluate their predictive power. An example of this strategy is a sustained three year study by Drexler *et al.* (2011) of emerging coronaviruses and astroviruses in a bat colony. Viral emergence and loads were assessed quantitatively over the entire study period. These data illuminated with fine-grained time resolution how bat colonies experienced and controlled viral infections. Another longitudinal study of viruses in a wild bat colony was presented by Linfa Wang (of Australia's Commonwealth Scientific and

**Box 2: Knowing the Baseline: the Need for Background Sampling**

A recurring theme at the workshop was the need to "know the baseline" of viruses occurring within their natural hosts, namely which viruses infect which host, vector and reservoir species at given places and times. Without such data, the significance of detecting a virus in an animal at any given place or time is uncertain. For example, detection of *Reston ebolavirus* (REBOV) in pigs made international news, and is significant both in its novelty and its unclear risk to human populations. However, it may be that pigs are historical hosts of REBOV globally, and that this detection is therefore not indicative of increased risk to humans. Without knowing the distribution of REBOV in pigs across space and time, the associated risk cannot be estimated.

Industrial Research Organisation, CSIRO) at the 2011 International Meeting for Emerging Diseases and Surveillance (IMED)[9]. For this study, weekly bat urine samples are collected and tested using PCR, viral culture and sequencing; study results are not yet published. While longitudinal sampling is more time consuming and labor intensive than sampling at one time point, it is essential for capturing the dynamics and variations in genetic signatures of viral evolutionary change.

Environmental background sampling for species jumps will require well designed and statistically supported plans that provide accurate estimations of $P_d$ from pre-event surveys. In contrast to post-event grab sampling (i.e., collecting one sample at one time), pre-event sampling methods will require utilization of designs developed for environmental contaminant and ecological surveys. These modalities include random, targeted, random grid, stratified, random stratified and transect methodologies (Box 3). $P_d$ can also be increased by the use of adaptive sampling strategies that adjust the sampling plan based on near real-time data obtained from field bioassays. Selection of the most appropriate approach should be driven by a thorough and prior assessment of the data quality objectives (DQO) for

## Box 3: Applying Environmental Sampling to Detection of Species Jumps: the Basics

To maximize the ***probability of detection*** ($P_d$) of a target (in this case, a virus) within a given area (e.g., multiple adjacent bat roosting sites), a site-specific sampling plan is required. Two key components of a sampling plan are ***sampling designs*** and ***data quality objectives*** (DQO).

**Sampling Designs:** There are numerous sampling designs appropriate for different sites, budgets and objectives. One can test the sensitivity of sampling designs using statistical tools like logistic regression to construct detection curves, which relate the sample numbers and the density of targets to $P_d$. This approach can help to reduce false negatives and increase the ability to detect rare viruses at low densities by enabling refinement of sampling strategies.

**Data Quality Objectives**: DQO are essentially standards that investigators set *prior to sampling* to insure that resulting data are *representative, sufficient and usable*. Some examples of DQO are:

- Sample numbers, collection points, analysis methods (e.g., viral genome sequencing)
- Sample documentation (e.g., photograph, video, manual logbook, barcode, GPS coordinates)
- Data management (e.g., electronic file format, spreadsheet, database, web-based platform)
- Quality assurance/quality control (QA/QC) (e.g., number of sample replicates, sample blanks)
- Statistical tools to measure data representativeness, sufficiency and usability (e.g., variance, sample size, standard error)

Well-developed DQO are required to determine whether a sampling design or site is worth extensive evaluation. As such, DQO figure heavily in cost-benefit analyses. For example, calculating the minimum number of samples required for representativeness (sample size) enables investigators to in turn calculate the minimum investment required.

For thorough descriptions of sampling designs and statistical analysis as well as DQO guidance, please see the Environmental Protection Agency's *Guidance on Choosing a Sampling Design for Environmental Data Collection (http://www.epa.gov/quality/qs-docs/g5s-final.pdf)*.

9 http://ww2.isid.org/Downloads/IMED2011_Presentations/IMED2011_Wang.pdf

the study (Box 3). Other important considerations for sampling plan success include communication and planning with the analytical laboratory (where samples will be processed); elicitation of subject matter experts; collaboration with scientists experienced in field sampling; and assembly of an expert panel to review the sampling approach and suggest corrections.

## PILOT STUDIES

Thus far we have proposed informatics and field sampling strategies to better understand how viral species jumps occur and better enable their prediction in future. Finally, we suggest pilot studies that will close gaps in SJOne as well as provide opportunities to exercise background sampling plans. Since emerging viruses with a history of human spillover may evolve to jump into human hosts, we see an opportunity to leverage and augment current research efforts to include collection of field and laboratory data to specifically determine differences between viral strains that exhibit host-specific differences in pathogenicity. Ripe for study are the genetic and phenotypic differences between Nipah and Nipah-like viruses in pig, bat and human hosts as well as differences between Ebola virus strains that are pathogenic or non-pathogenic in humans. Finally, we discuss the prospect of horizontal gene transfer as a key enabler of species jumps.

***Nipah and Nipah-like Viruses***

Henipaviruses are an important group of emerging zoonotic RNA viruses whose primary hosts are large fruit bats of the genus *Pteropus*. Henipaviruses include Hendra and Nipah species, which have been isolated from bats in Australia, Asia, Africa, and Madagascar (Hayman *et al*. 2008; Epstein *et al*. 2008). Hendra viruses from Australian fruit bats have caused small, isolated fatal disease outbreaks among horses and fatal spillover cases among humans exposed to infected horses (Plowright *et al*. 2011), while Nipah viruses from

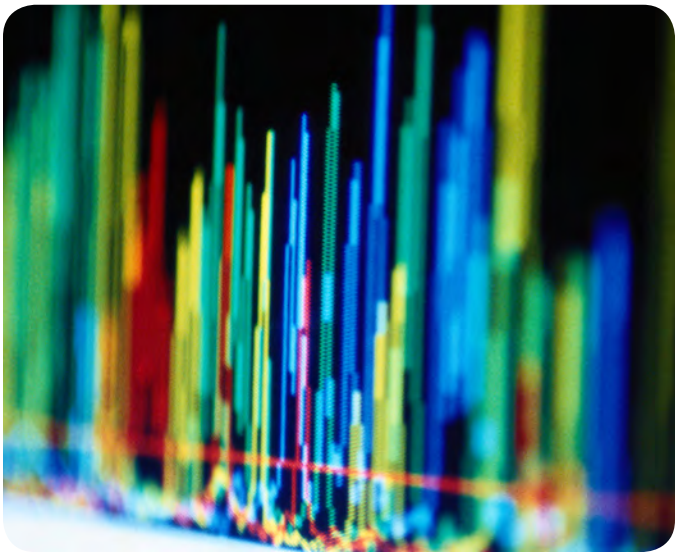
the Asian mainland have caused major outbreaks involving pigs, humans, horses, cats, dogs, goats, and chickens (Bunning 2001). Although human-to-human transmission of Nipah virus was never reported from Malaysia and Singapore, where fatal human cases were first observed in conjunction with mass outbreaks in domestic pigs (Goh *et al*. 2000; Chew *et al*. 2000), a number of subsequent Nipah outbreaks in Bangladesh and eastern India have been associated with exposure to bat-contaminated palm tree sap and iterative human-to-human transmission events (Chadha *et al*. 2006; Gurley *et al*. 2007; Luby *et al*. 2009a, 2009b). As-yet unnamed Henipahviruses have been detected in fruit bats in northern India, West Africa and Madagascar, although no species jumps or spillover events have been detected in humans or domesticated animal species (Iehlé *et al*. 2007; Hayman *et al*. 2008; Epstein *et al*. 2008).

Detailed analyses of genomic differences between Nipah virus strains from Malaysia and Singapore

and those from Bangladesh and India could be designed to identify the genetic basis for human-to-human transmission capability in Nipah viruses, because human-to-human transmission of Nipah virus has occurred in India and Bangladesh, but has not been reported from human outbreaks in Malaysia and Singapore. (This assumes that such differences in transmission to humans are not due to uneven reporting of human infections or human host genetic variation.) Such analyses could be performed using laboratory data (e.g., characterization of variant viral proteins and their functions in host cells *in vitro*) as well as field data (e.g., sequences from longitudinally collected samples). Of greatest impact would be studies that are designed to fill extant data gaps and collect data that can be shared with other stakeholders via platforms like SJOne.

***Ebola Viruses***

There are five known species of Ebola virus: Bundibugyo, Cote d'Ivoire, Reston, Sudan and Zaire. The relative virulence of these viruses varies greatly among infected humans and other mammalian species. *Zaire* and *Sudan ebolaviruses* are associated with mortality rates of 50-90% in humans, while *Côte d'Ivoire ebolavirus* has only been reported from a non-fatal human case and Reston ebolavirus is only associated with asymptomatic infections in humans (Towner *et al*. 2008; Bausch 2011).



*Reston ebolavirus* causes asymptomatic infections in domestic pigs, while *Zaire ebolavirus* has been associated with fatal infections of wild African pigs (Lahm *et al*. 2007) and shown experimentally to cause symptomatic disease associated with pig-to-pig transmission in domestic swine (Kobinger *et al*. 2011). The natural host reservoirs for these viruses are currently unknown, although there is evidence that certain species of African fruit bats may serve as reservoirs (Biek *et al*. 2006; Porrut *et al*. 2009). The closely-related Marburg virus has been recovered from the *Rousettus aegyptiacus* fruit bat, and human spillover cases have been linked to exposure to bats in caves where infected bats were recovered (Towner *et al*. 2009). Detailed analyses of genomic differences among Ebola virus species could reveal clues as to why certain viruses can infect multiple host species without causing disease in all of them. Much remains unknown about the mechanisms these viruses use to infect hosts. Only very recently was a human host receptor identified for *Zaire ebolavirus* and *Lake Victoria Marburg virus* (Kondratowicz *et al*. 2011) – despite a decades-long history of human spillover.

***Horizontal Gene Transfer***

Genetic changes resulting from the acquisition of new genes or sequences can be an important factor in viral species jumps. Horizontal gene transfer (HGT) is a type of gene flow that occurs in bacteria and viruses, which can produce new strains with enhanced virulence and the ability to infect new hosts. Although HGT occurs more frequently in bacteria, several families of viruses (including Poxvirus and Herpesvirus species) are known to have acquired genes from hosts through HGT (Fu *et al*. 2008; Monier *et al*. 2009; Odom *et al*. 2009). HGT may occur between host cells and viruses as well as between viruses themselves (Liu *et al*. 2010). The incorporation of host genes into viral genomes may facilitate transmission and modulate host immune response mechanisms (Odom *et al*. 2009). Recent research suggests that HGT between double-stranded RNA viruses and non-bacterial hosts occurs quite frequently, and may give rise to

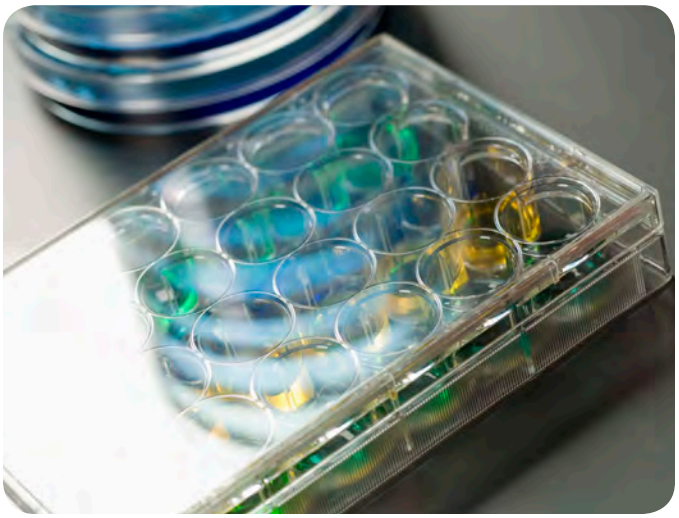
functionally important new genes in host genomes that play a role in host evolution (Liu *et al*. 2010).

Workshop participant David Krakauer, Professor and Chair of the Faculty at the Santa Fe Institute, cited evidence of HGT among bacterial species (Ochman *et al*. 2000) as examples of ancient species jumps. Monitoring genomic change over time will reveal evolutionary changes in viral strains, but if species jumps are made possible by HGT (e.g., gene exchange between co-infecting viruses in a single host), then these abrupt changes might not be foreseen in timely fashion. Theoretical work presented by Michael Deem, Professor at Rice University, reinforced the importance of HGT as a major mechanism for generating viral protein diversity (Bogarad & Deem 1999). However, the frequency and significance of HGT among emerging viruses are understudied, so it is difficult to determine the contributions that HGT make to viral species jumps. Krakauer presented a "core and satellite" model for viruses, in which a highly conserved genetic core is orbited by less conserved satellite genes, with the totality being equal to the viral genome. Krakauer and colleagues compared core and satellite genes from four DNA virus families and found that viruses with smaller genomes are more promiscuous in their ability to acquire satellite genes from other viral species. This analysis also found that these viruses were more likely to gain satellite genes than to lose them (Zanotto & Krakauer 2008).

If it is true that species jumps may result from gene gain, and that viruses with smaller genomes are more likely to acquire genes, then smaller viruses may be more likely to jump hosts. However, it is important to note that this analysis was performed using DNA viruses; it remains to be seen whether the correlation between genome size and promiscuity holds for RNA viruses. Holmes and Rambaut (2004), while acknowledging that coronaviruses are among the few RNA viruses known to undergo HGT, argue against HGT as a mechanism for the SARS coronavirus species jump to humans. They and others (Pepin *et al.* 2010) point out that the available sample size is too small to draw conclusions about the evolutionary history of SARS coronavirus, and suggest that biosurveillance systems tailored to recognize salient changes in viral fitness could cue early warning of species jumps like the one behind the 2003 SARS outbreak. We therefore propose additional coronavirus pilot studies, such as those described by Drexler *et al.* (2011), be expanded in scope beyond PCR assessment of viral load, to

include complete genomic analysis of how bat colony coronavirus isolates evolve over space and time.

It is unclear whether gene gains could be detected in a timely fashion – that is, before a virus jumps and causes outbreaks in a new host species. In this case, surveillance for gene gains may not enable prediction, but could be used for early detection and mitigation. For example, if gene gains were detected early in an outbreak, then more rigorous mitigation strategies could be implemented to prevent transmission and ward off a potential species jump. A recent study has demonstrated that longitudinal surveys can detect HGT, as well as rare viral sequences (Vijaykrishna *et al.* 2011). A 12-year surveillance study of influenza viruses in Hong Kong swine found evidence of one HGT that occurred repeatedly, suggesting that this HGT is not random but a response to evolutionary pressure on the virus. This study also demonstrates that longitudinal surveys enable detection of rare viral sequences – known to be rare because they occur among a large number of total samples collected. Detection of rare sequences further enabled investigators to hypothesize that they were poorly adapted for large-scale spread, because they "died out" over time. Ultimately, more data from longitudinal studies of different virus-host systems will bolster the framework of understanding for how species jumps occur.



## CONCLUSION

Currently, most zoonotic surveillance efforts seek ecological drivers of spillover events – factors like climate, land use and population migrations driving infections that do not spread between humans. By contrast, we focus here on the evolutionary drivers of species jumps – the genetic changes over time driving infections that spread efficiently among humans. We see an opportunity to use field surveillance and laboratory data to improve understanding of how viral species jumps occur, and have suggested three approaches to build a mechanistic framework of understanding: by curating extant data and fusing them with new and existing bioinformatics tools in SJOne; by generating new field sampling plans informed by environmental contaminant and ecological surveys; and by pursuing pilot studies likely to yield new data of direct relevance to species jumps. With a mechanistic framework of understanding bolstered by sufficient quantities of relevant data, more sophisticated models can be built, and the risks posed by viral species jumps can be calculated, managed and someday predicted.

## ACKNOWLEDGEMENTS

The authors thank Hsinchun Chen, Helen Cui, Michael Deem, Lucky Gunasakara, David Krakauer, Richard Scheuermann, James Lloyd-Smith, Juliet Pulliam, Stephan Velsko and John Yin for sharing their insights and expertise at the workshop. We also thank Sarah Cobey, Greg Glass and Peter Daszak for expertly moderating the workshop sessions.

# REFERENCES


Bausch DG. (2011) Ebola Virus as a Foodborne Pathogen? Cause for Consideration, but Not Panic J Infect Dis. Epubl ahead of print May 12, 2011.
http://www.oxfordjournals.org/our_journals/jid/jir201.pdf

Biek R, Walsh PD, Leroy EM, Real LA. (2006) Recent Common Ancestry of Ebola Zaire Virus Found in a Bat Reservoir. PLoS Pathog 2(10): e90.
http://www.plospathogens.org/article/info%3Adoi%2F10.1371%2Fjournal.ppat.0020090

Bogarad LD and Deem MW (1999) A hierarchical approach to protein molecular evolution.PNAS 96(6): 2591-2595.
http://www.pnas.org/content/96/6/2591.abstract?sid=db987c09-67d6-4526-96c8-cc6182080472

Bunning M. (2001) Nipah virus outbreak in Malaysia, 1998-1999. J Swine Health Prod 9(6): 295-299.
http://www.aasv.org/shap/issues/v9n6/v9n6p295.pdf

Cavalli-Sforza LL, Feldman MW. (2003) The application of molecular genetic approaches to the study of human evolution. Nature Genetics 33, 266 - 275 (2003).
http://www.nature.com/ng/journal/v33/n3s/full/ng1113.html

Chadha MS, Comer JA, Lowe L, Rota PA, Rollin PE, Bellini WJ, Ksiazek TG, Mishra AC. (2006) Nipah virus-associated encephalitis outbreak, Siliguri, India. Emerg Infect Dis 12: 235-240.
http://www.cdc.gov/ncidod/EID/vol12no02/05-1247.htm

Chew MH, Arguin PM, Shay DK, Goh KT, Rollin PE, Shieh WJ, Zaki SR, Rota PA, Ling AE, Ksiazek TG, Chew SK, Anderson LJ. (2000) Risk factors for Nipah virus infection among abattoir workers in Singapore. J Infect Dis 181(5): 1760-1763.
http://www.journals.uchicago.edu/doi/pdf/10.1086/315443

Drexler JF, Corman VM, Wegner T, Tateno AF, Zerbinati RM, Gloza-Rausch F, Seebens A, Müller MA, Drosten C. (2011) Amplification of Emerging Viruses in a Bat Colony. Emerg Infect Dis 17(3) 2011 Mar.
http://www.cdc.gov/eid/content/17/3/449.htm

Epstein JH, Prakash V, Smith CS, Daszak P, McLaughlin AB, Meehan G, Field HE, Cunningham AA. (2008) Henipavirus infection in fruit bats (Pteropus giganteus), India. Emerg Infect Dis 14(8): 1309-1311.
http://www.cdc.gov/eid/content/14/8/1309.htm

Fu M, Deng R, Wang J, Wang X. (2008) Detection and analysis of horizontal gene transfer in herpesvirus. Detection and analysis of horizontal gene transfer in herpesvirus. Virus Res 131(1): 65-76.
http://www.sciencedirect.com/science/article/pii/S016817020700322X

Goh KJ, Tan CT, Chew NK, Tan PS, Kamarulzaman A, Sarji SA, Wong KT, Abdullah BJ, Chua KB, Lam SK. (2000) Clinical features of Nipah virus encephalitis among pig farmers in Malaysia. N Engl J Med 342(17): 1229-1235.
http://www.nejm.org/doi/full/10.1056/NEJM200004273421701

Gurley ES, Montgomery JM, Hossain MJ, Bell M, Azad AK, Islam MR, Molla MA, Carroll DS, Ksiazek TG, Rota PA, Lowe L, Comer JA, Rollin P, Czub M, Grolla A, Feldmann H, Luby SP, Woodward JL, Breiman RF. (2007) Person-to-person transmission of Nipah virus in a Bangladeshi community. Emerg Infect Dis 13: 1031-1037.
http://www.cdc.gov/eid/content/13/7/1031.htm

Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, Luo SW, Li PH, Zhang LJ, Guan YJ, Butt KM, Wong KL, Chan KW, Lim W, Shortridge KF, Yuen KY, Peiris JS, Poon LL. (2003) Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302: 276-278.
http://www.sciencemag.org/content/302/5643/276.long

Hayman DT, Suu-Ire R, Breed AC, McEachern JA, Wang L, Wood JL, Cunningham AA. (2008) Evidence of henipavirus infection in West African fruit bats. PLoS One 3(7): e2739.
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0002739

Holmes EC, Rambaut A. (2004) Viral evolution and the emergence of SARS coronavirus. Phil. Trans. R. Soc. Lond. B 359:1059-106.
http://rstb.royalsocietypublishing.org/content/359/1447/1059.full.pdf

Iehlé C, Razafitrimo G, Razainirina J, Andriaholinirina N, Goodman SM, Faure C, Georges-Courbot MC, Rousset D, Reynes JM. (2007) Henipavirus and Tioman virus antibodies in pteropodid bats, Madagascar. Emerg Infect Dis 13(1): 159-61.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2725826/pdf/06-0791.pdf

Institute of Medicine (2007) Review of the DoD-GEIS Influenza Programs: Strengthening Global Surveillance and Response.
http://www.nap.edu/catalog.php?record_id=11974

Institute of Medicine Consensus Report (2009) Sustaining Global Surveillance and Response to Emerging Zoonotic Diseases.
http://www.iom.edu/Reports/2009/ZoonoticDisease.aspx

Jensen LJ, Saric J, Bork P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet. 7(2): 119-29.
http://www.nature.com/nrg/journal/v7/n2/abs/nrg1768.html

Kobinger GP, Leung A, Neufeld J, Richardson JS, Falzarano D, Smith G, Tierney K, Patel A, Weingartl HM. (2011) Replication, Pathogenicity, Shedding, and Transmission of Zaire ebolavirus in Pigs. J Infect Dis. 2011 May 12. [Epub ahead of print].
http://jid.oxfordjournals.org/content/early/2011/05/12/infdis.jir077.abstract

Kondratowicz AS, Lennemann NJ, Sinn PL, Davey RA, Hunt CL, Moller-Tank S, Meyerholz DK, Rennert P, Mullins RF, Brindley M, Sandersfeld LM, Quinn K, Weller M, McCray, Jr. PB, Chiorini J, Maury W. (2011) T-cell immunoglobulin and mucin domain 1 (TIM-1) is a receptor for Zaire Ebolavirus and Lake Victoria Marburgvirus. PNAS 108(20): 8426-8431.
http://www.pnas.org/content/108/20/8426

Lahm SA, Kombila M, Swanepoel R, Barnes RF. 2007. Morbidity and mortality of wild animals in relation to outbreaks of Ebola haemorrhagic fever in Gabon, 1994-2003. Trans R Soc Trop Med Hyg 101(1): 64-78.
http://www.tropicalmedandhygienejrnl.net/article/S0035-9203(06)00212-4/abstract

Lim K-i, Lang T, Lam V, Yin J (2006) Model-Based Design of Growth-Attenuated Viruses. PLoS Comput Biol 2(9): e116. doi:10.1371/journal.pcbi.0020116.
http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.0020116

Liu H, Fu Y, Jiang D, Li G, Xie J, Cheng J, Peng Y, Ghabrial SA, Yi X. (2010) Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. J Virol. 84(22): 11876-11887.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2977895/?tool=pubmed

Luby SP, Gurley ES, Hossain MJ. (2009a). Transmission of human infection with Nipah virus. Clin Infect Dis 49(11): 1743-1748.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2784122/?tool=pubmed

Luby SP, Hossain MJ, Gurley ES, Ahmed BN, Banu S, Khan SU, Homaira N, Rota PA, Rollin PE, Comer JA, Kenah E, Ksiazek TG, Rahman M. (2009b) Recurrent zoonotic transmission of Nipah virus into humans, Bangladesh, 2001–2007. Emerg Infect Dis 15: 1229-1235.
http://www.cdc.gov/EID/content/15/8/1229.htm

Monier A, Pagarete A, de Vargas C, Allen MJ, Read B, Claverie JM, Ogata H. (2009) Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. Genome Res. 2009 19: 1441-1449.
http://genome.cshlp.org/content/19/8/1441.full

Muller H-M, Kenny EE, Sternberg PW. (2004) Textpresso: An ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2(11): e309.
http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0020309

Nugent G. (2011) Maintenance, spillover and spillback transmission of bovine tuberculosis in multi-host wildlife complexes: a New Zealand case study. Vet Microbiol 2011 Feb 24. Epub ahead of print.
http://www.sciencedirect.com/science/article/B6TD6-5281SNC-5/2/158991686e565fd352a62992790c39af

Ochman H, Lawrence JG, Groisman EA. (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299-304.
http://www.biochem.arizona.edu/ochman/Papers/Ochman_Nature2000.pdf

Odom MR, Hendrickson RC, Lefkowitz EJ. (2009) Poxvirus protein evolution: family wide assessment of possible horizontal gene transfer events. Virus Res 144(1-2): 233-249.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2779260/?tool=pubmed

Palmer MV. (2007) Tuberculosis: a reemerging disease at the interface of domestic animals and wildlife. Curr Top Microbiol Immun 315:195-215.
http://www.springerlink.com/content/m2p52451u945nh48/fulltext.pdf

Parrish CR, Holmes EC, Morens DM, Park E-C, Burke DS, Calisher CH, Laughlin CA, Saif LJ, Daszak P. (2008) Cross-Species Virus Transmission and the Emergence of New Epidemic Diseases. Micro Mol Biol Rev 72(3):457-470.
http://mmbr.asm.org/cgi/content/abstract/72/3/457

Pepin KM, Lass S, Pulliam JRC, Read AF, Lloyd-Smith JO. (2010) Identifying genetic markers of adaptation for surveillance of viral host jumps. Nature Rev Microbio 8: 802-813.
http://www.nature.com/nrmicro/journal/v8/n11/full/nrmicro2440.html

Plowright RK, Foley P, Field HE, Dobson AP, Foley JE, Eby P, Daszak P. (2011) Urban habituation, ecological connectivity and epidemic dampening: the emergence of Hendra virus from flying foxes (Pteropus spp.). Proc Biol Sci. 2011 May 11. [Epub ahead of print].
http://rspb.royalsocietypublishing.org/content/early/2011/05/06/rspb.2011.0522.full.pdf+html

Pourrut X, Souris M, Towner JS, Rollin PE, Nichol ST, Gonzalez JP, Leroy E. (2009) Large serological survey showing cocirculation of Ebola and Marburg viruses in Gabonese bat populations, and a high seroprevalence of both viruses in Rousettus aegyptiacus. BMC Infect Dis 9: 159.
http://www.biomedcentral.com/1471-2334/9/159

Prakash M, (2008) Molecular Biology of Evolution, Discovery Publishing House Pvt Ltd, New Delhi, p. 191.
http://books.google.com/books?id=6x2UmpU6grsC&pg=PA191&dq=%22evolutionary+driver%22&hl=en&ei=NyzmTZj0LsTr0gHy9PD3Cg&sa=X&oi=book_result&ct=result&resnum=5&sqi=2&ved=0CD8Q6AEwBA#v=onepage&q=%22evolutionary%20driver%22&f=false

Reichman OJ, Jones MB, Schildhauer MP (2011) Challenges and Opportunities of Open Data in Ecology. Science 331(6018): 703-705.
http://www.sciencemag.org/content/331/6018/703.abstract?sid=cbb7dad9-45de-4b65-b216-0321471665f7

Russell KL, Rubenstein J, Burke RL, Vest KG, Johns MC, Sanchez JL, Meyer W, Fukuda MM, Blazes DL (2011) The Global Emerging Infection Surveillance and Response System (GEIS), a U.S. government tool for improved global biosurveillance: a review of 2009. BMC Public Health 11(Suppl 2): S2.
http://www.biomedcentral.com/1471-2458/11/S2/S2

Sueker JJ, Blazes DL, Johns MC, Blair PJ, Sjoberg PA, Tjaden,JA, Montgomery JM, Pavlin JA, Schnabel DC, Eick AA, Tobias S, Quintana M, Vest KG, Burke RL, Lindler LE, Mansfield JL, Erickson RL, Russell KL, Sancheza JL for the DoD Influenza Working Group. (2010) Influenza and respiratory disease surveillance: the US military's global laboratory-based network. Influenza and Other Respiratory Viruses 4(3), 155–161.
http://www.ncbi.nlm.nih.gov/pubmed/20409212

Towner JS, Amman BR, Sealy TK, Carroll SA, Comer JA, Kemp A, Swanepoel R, Paddock CD, Balinandi S, Khristova ML, Formenty PB, Albarino CG, Miller DM, Reed ZD, Kayiwa JT, Mills JN, Cannon DL, Greer PW, Byaruhanga E, Farnon EC, Atimnedi P, Okware S, Katongole-Mbidde E, Downing R, Tappero JW, Zaki SR, Ksiazek TG, Nichol ST, Rollin PE. (2009) Isolation of genetically diverse Marburg viruses from Egyptian fruit bats. Isolation of genetically diverse Marburg viruses from Egyptian fruit bats. PLoS Pathog 5(7): e1000536.
http://www.plospathogens.org/article/info%3Adoi%2F10.1371%2Fjournal.ppat.1000536

Towner JS, Sealy TK, Khristova ML, Albariño CG, Conlan S, Reeder SA, Quan PL, Lipkin WI, Downing R, Tappero JW, Okware S, Lutwama J, Bakamutumaho B, Kayiwa J, Comer JA, Rollin PE, Ksiazek TG, Nichol ST. (2008) Newly discovered Ebola virus associated with hemorrhagic fever outbreak in Uganda. PLoS Pathog 4(11): e1000212.
http://www.plospathogens.org/article/info%3Adoi%2F10.1371%2Fjournal.ppat.1000212

Vijaykrishna D, Smith GJD, Pybus OG, Zhu H, Bhatt S, Poon LLM, Riley S, Bahl J, Ma SK, Cheung CL, Perera RAPM, Chen H, Shortridge KF, Webby RJ, Webster RG, Guan Y, Peiris JSM. (2011) Long-term evolution and transmission dynamics of swine influenza A virus. Nature 473: 519-522.
http://www.nature.com/nature/journal/v473/n7348/full/nature10004.html

Wang LF, Eaton BT. (2007) Bats, civets and the emergence of SARS. Curr Top Microbiol Immunol 315: 325-344.
http://www.springerlink.com/content/u522k72mml2043p8/

Zanotto PMdA, Krakauer DC. (2008) Complete Genome Viral Phylogenies Suggests the Concerted Evolution of Regulatory Cores and Accessory Satellites. PLoS ONE 3(10): e3500.
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0003500

# APPENDIX 1: Annotated List of Genomic Databases and Analytical Resources

**BioPortal** http://ai.arizona.edu/research/bioportal/
This database contains a phylogenetic analysis module for pathogen DNA to determine the genetic relationship between various strains, and identify possible sources or mutation. The module also provides color-coded analyses of outbreak occurrences based on distance in genetic space.

**BSORF Bacillus subtilis Open Reading Frames** http://bacillus.genome.jp/
Genomics database for *Bacillis subtilis* operated by Kyoto University in Japan.

**CMR Comprehensive Microbial Resource** http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi
The Comprehensive Microbial Resource (CMR) is an open-access website with information on publicly available, complete prokaryotic genomes. In addition to the convenience of having all of the organisms on a single website, common data types across all genomes in the CMR make searches more meaningful, and cross genome analysis highlight differences and similarities between the genomes.

**CoVDB** http://covdb.microbiology.hku.hk/
Database hosted by the University of Hong Kong which provides genomic data for 268 coronaviruses isolated from humans, mammals and birds.

**DDBJ DNA Data Bank of Japan** http://www.ddbj.nig.ac.jp/Welcome.html.en
DNA Data Bank of Japan is the sole nucleotide sequence data bank in Asia, which is officially certified to collect nucleotide sequences from researchers and to issue the internationally recognized accession number to data submitters. Because DDBJ exchanges collected data with EMBL-Bank/EBI and GenBank/NCBI on a daily basis, these three data banks through a virtually integrated database called the "International Nucleotide Sequence Database (INSD)".

**DPVweb** http://www.dpvweb.net/
DPVweb a sequence feature database (DPVweb) that contains all sequences of viruses, viroids and satellites of plants, fungi and protozoa, that are complete or which encode one or more gene using accession numbers from EMBL and GenBank.

**EcoGene** http://www.ecogene.org/
*E. coli* database collection operated by the University of Miami.

**EMBL Database** http://www.ebi.ac.uk/genomes/index.html
The EMBL database is operated by the European Bioinformatics Institute (EBI), a non-profit academic organization affiliate of the European Molecular Biology Laboratory (EMBL). As of 21 Feb 2011, the EMBL database contains genome data for 2393 viruses, 1440 bacteria, and 53 viroids.

**Ensembl** http://www.ensembl.org/
The Ensembl project produces open-access online genome databases for vertebrates and other eukaryotic taxa. Six databases sites are now available: Ensembl (vertebrate genomes), Ensembl Bacteria, Ensembl Protists, Ensembl Metazoa, Ensembl Plants and Ensembl Fungi.

**EuPathDB** http://eupathdb.org/eupathdb/
EuPathDB Bioinformatics Resource Center for Biodefense and Emerging/Re-emerging Infectious Diseases is a portal for accessing genomic-scale datasets associated with the eukaryotic pathogens (Cryptosporidium, Encephalitozoon, Entamoeba, Enterocytozoon, Giardia, Leishmania, Neospora, Plasmodium, Toxoplasma, Trichomonas and Trypanosoma).

**European Hepatitis C Virus Database [euHCVDB]** http://euhcvdb.ibcp.fr
The euHCVdb is an extension of the French HCV Database that is updated on a monthly basis from the EMBL nucleotide sequence database and maintained in the PostgreSQL relational database management system (RDMS).

**European Nucleotide Archive** http://www.ebi.ac.uk/ena/
The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.

**Gemina** http://gemina.igs.umaryland.edu/cgi-bin/gemina/MakeFrontPages.cgi

Gemina is a web-based system designed to identify infectious pathogens and their representative genomic sequences through selection of associated epidemiology metadata. Gemina supports the development of DNA signature-based assays for the detection of pathogens or sets of pathogen through the Insignia Signature Pipeline at the University of Maryland. View the Quick Start or the Gemina Tutorial for help on searching the database.

**GenBank** http://www.ncbi.nlm.nih.gov/genbank/index.html

GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. GenBank is a comprehensive database that contains publicly available nucleotide sequences for more than 380,000 organisms classified at the genus taxon or below, obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects. GenBank is built and distributed by the National Center for Biotechnology Information of the US National Institutes of Health (NIH).

**GenoList PF4 Genomic Integration and Analysis Platform** http://genolist.pasteur.fr/GenoList

GenoList is an integrated environment dedicated to querying and analyzing genome data from bacterial species operated by the Pasteur Institute in France.

**Genome Encyclopedia of Microbes GEM** http://www.gem.re.kr/

A genomic resource and database portal supported by Ministry of Science and Technology, Republic of Korea.

**Genome Portal** http://genome.jgi-psf.org/

Open-access online genomic sequence database operated by the Department of Energy - Joint Genome Institute that contains genomic sequence data on fungi, plants, microbes, and organisms.

**GIBV Genome Information Broker for Viruses** http://gib-v.genes.nig.ac.jp/

Genome Information Broker for Viruses (GIB-V) is a complete virus genome data repository. We extracted 79,297 complete virus genomes or segments data comprehensively from International Nucleotide Sequence Database Collaboration (INSDC; DDBJ, EMBL database and GenBank) and stored in this system.

**GISAID EpiFlu** http://platform.gisaid.org/

Featuring the world's most complete collection of influenza sequences containing associated metadata, both clinical & epidemiological. Its functionality continues to be tailored to the needs of influenza researchers from both the human and the veterinary fields. Conceptualized by the knowledge of influenza scientists, it is powered by developers at the Max-Planck-Institute for Informatics and a3systems GmbH, Saarbrücken, Germany.

**GOLD Genome Online Database** http://genomesonline.org/

Open-access cooperatively managed site started in 1997 that provides access to data on 10,000 genomes including more than 1,000 microbial genomes in Excel spreadshett format.

**HCV Database** http://hcv.lanl.gov/

The Hepatitis C Virus (HCV) sequence database is a web-accessible sequence database of annotated HCV-associated genetic data operated by the Los Alamos National Laboratory.

**HCVDB - Hepatitis C Virus Database** http://hepatitis.ibcp.fr/

**HIV Databases** http://www.hiv.lanl.gov/content/index

The HIV databases operated by the Los Alamos National Laboratory contain data on HIV genetic sequences, immunological epitopes, drug resistance-associated mutations, and vaccine trials.

**ICTVdb** http://www.ictvdb.org/index.htm

The ICTVdB contains an taxonomic index of viruses and list of approved virus names linked to virus descriptions coded from information in Virus Taxonomy which also incorporates the plant virus database VIDEdB. ICTVdb provides searchable descriptions of virus isolates, species, genera, families, orders, virus images of many viruses, and links to genomic and protein databanks. Originally developed at the Australian National University (ANU) with support of US National Science Foundation (NSF) under sponsorship by the American Type Culture Collection (ATCC).

**INSD International Nucleotide Sequence Database** http://www.insdc.org/

The International Nucleotide Sequence Databases (INSD) contains shared data from DDBJ, European Nucleotide Archive, and GenBank.

**IRD Influenza Research Database** http://www.fludb.org/brc/statsAutomation.do?decorator=influenza

Open-access online database with influenza segment and protein sequences, avian and non-human mammalian surveillance data, virus phenotypic characteristics, influenza strain information, and immune epitope data. Operated by a public-private consortium that includes University of Texas- Southwestern Medical Center, NIH/NIAID, University of California – Davis, Los Alamos National Laboratory, and Northrup Grummond (among others).

**Influenza Sequence and Epitope Database (ISED)** http://influenza.korea.ac.kr/ISED2/index_3.jsp

ISED catalogues the influenza sequence and epitope information obtained in countries worldwide and currently hosts a total of 50402 influenza A and 5215 influenza B virus sequence data including pandemic A/H1N1 2009 virus sequences collected in 42 countries, and a total of 545 amantadine-resistant influenza virus sequences collected in Korea.

**Integrated Microbial Genomes (IMG)** http://img.jgi.doe.gov/cgi-bin/pub/main.cgi

IMG provides an open-access online resource for comparative analysis and annotation of publicly available genomes. Plasmids that are not part of a specific microbial genome sequencing project and phage genomes are also included into IMG in order to increase its genomic context for comparative analysis.

**IVDB - Influenza Virus Database** http://influenza.psych.ac.cn/

DB hosts complete genome sequences of influenza A virus generated by Beijing Institute of Genomics and annotated versions of other published influenza virus sequences. Our Nucleotide sequences are ranked into 7 categories according to sequence content and integrity. IVDB provides a series of tools and viewers for comparative analysis of the viral genomes, genes, genetic polymorphisms and phylogenetic relationships. Database access site is hosted by the Chinese Academy of Sciences - Institute of Psychology.

**Department of Energy - Joint Genome Institute (JGI)** http://www.jgi.doe.gov

The U.S. Department of Energy Joint Genome Institute is a cooperative project between five national laboratories (Lawrence Berkeley N.L., Lawrence Livermore N.L., Los Alamos N.L., Oak Ridge N.L., Pacific Northwest N.L.), and the HudsonAlpha Institute for Biotechnology that provides genomic sequence data on fungi, plants, microbes, and organisms through an open-access online Genome Portal. JGI database includes 770 Prokaryotic microbial genomes, of which 525 are complete.

**Kyoto Encyclopedia of Genes and Genomes (KEGG)** http://www.genome.jp/kegg-bin/get_htext?Viruses+-e

KEGG GENOME has been a collection of organisms with known complete genome sequences supplemented by those with massive EST datasets. It now contains metagenomes representing environmental samples (ecosystems) of genome sequences for multiple species. In future, virus genomes will be integrated and virus genes annotated.

**Microbial Genome Database for Comparative Analysis** http://mbgd.genome.ad.jp/

MBGD is a database for comparative analysis of completely sequenced microbial genomes, the number of which is now growing rapidly. The aim of MBGD is to facilitate comparative genomics from various points of view such as ortholog identification, paralog clustering, motif analysis and gene order comparison.

**Microbial Rosetta Stone** http://www.microbialrosettastone.com/

The Microbial Rosetta Stone (MRS) is a database that relates microorganism names, taxonomic classifications, diseases, and scientific literature for the most important human, animal and plant microbial pathogens, with linkage to public genomic sequence databases. The database was created as a resource for biosecurity researchers in government, academia or industry who are not expert microbiologists, but have a research interest in infectious microbes.

**MiMi Web** http://mimi.ncibi.org/MimiWeb/main-page.jsp

MiMi Web gives you an easy to use interface to a rich NCIBI data repository for conducting your systems biology analyses. This repository includes the MiMI database, PubMed resources updated nightly, and text mined from biomedical research literature. The MiMI database comprehensively includes protein interaction information that has been integrated and merged from diverse protein interaction databases and other biological sources. With MiMI, you get one point of entry for querying, exploring, and analyzing all these data.

**NCBI National Center for Biotechnology Information Databases** http://www.ncbi.nlm.nih.gov/

NCBI's sequence databases provide access to genome data from sequencing projects from around the world through GenBank and its subsidiary database archives (Entrez Genomes, EST, GSS, HTG, SNPs).

**NCBI Entrez Genome / Viral Genomes** http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239

The NCBI Viral Genomes Resource currently contains 3764 Reference Sequences for 2567 viral genomes and 41 Reference Sequences for viroids, and about 6,500 additional complete sequences for more than 1,500 different viral species. Whenever more than one complete genome for a species is available, a pairwise global alignment of a reference sequence with each additional sequence is provided.

**Next Generation Biology Workbench** http://www.ngbw.org/

The Next Generation Biology Workbench is a free resource for research and education in Bioinformatics, Genomics, Proteomics, and Phylogenetics. The NGBW is a re-engineering of the Biology Workbench which was designed by Shankar Subramaniam and his group to provide an integrated environment where tools, user data, and public data resources can be easily accessed.

**PAIDB Pathogenicity Island Database** http://www.gem.re.kr/paidb

PAIDB is a comprehensive relational database of all the reported pathogenicity islands (PAIs) and potential PAI regions which were predicted by a method that combines feature-based analysis and similarity-based analysis. PAIs are genetic elements whose products are essential to the process of disease development that can be horizontally (laterally) transferred from other microbes and are important in evolution of pathogenesis.

**PhEVER** http://pbil.univ-lyon1.fr/databases/phever/index.php

PhEVER [Phylogenetic Exploration of Viruses Evolutionary Relationships] is an open-access genomic database operated by the University of Lyon (France). The database contains pre-computed alignment and phylogenies for homologous gene families of (i) sequences from different viruses and (ii) viral sequences and sequences from cellular organisms. Provides completely sequenced genome data on 2426 non-redundant viral genomes, 1007 non-redundant prokaryotic genomes, and 43 eukaryotic genomes ranging from plants to vertebrates. Provides clustering of proteins into homologous families containing at least one viral sequence, as well as alignments and phylogenies for each of these families.

**PathogenPortal** http://www.pathogenportal.org/portal/portal/PathPort/Home

Pathogen Portal is a repository linking to the Bioinformatics Resource Centers (BRCs) sponsored by the National Institute of Allergy and Infectious Diseases (NIAID) and maintained by The Virginia Bioinformatics Institute.

**RNA Virus Database** http://virus.zoo.ox.ac.uk/rnavirusdb/

Relational database for RNA viruses containing 1062 virus genomes, including dsRNA viruses, Retro-transcribing viruses, ssRNA negative-strand viruses, and ssRNA positive-strand viruses operated by the University of Oxford (UK), University of Edinburgh (UK), University of Aukland (NZ), and the University of Kwa-Zulu Natal (RSA).

**Subviral RNA Database** http://subviral.med.uottawa.ca/cgi-bin/home.cgi

The Subviral RNA database is a web-based environment that facilitates the research and analysis of viroids, satellite RNAs, satellites viruses, the human hepatitis delta virus, and related RNA sequences operated by the University of Ottawa (Canada). In addition to 2877 genetic sequences tehemselves, data entries include links to the original GenBank, EMBL and Medline records and may include supplemental data on position and secondary structures of the self-catalytic RNAs, prediction of the most stable secondary structures, multiple sequence aligments, duplicated sequences, etc.

**T4-like Genome Database** http://phage.ggc.edu/

The T4-like bacteriophage genome database is a compilation of information resources for T4-like Bacteriophage Genomes includes GenBank data and unpublished and published data from the Tulane T4-like Sequencing project and outside researchers currently hosted by Georgia Gwinnett College (US).

**UCSC Genome Bioinformatics Site** http://genome.ucsc.edu/

This site operated by the University of California Santa Cruz (US) contains the reference sequences and working draft assemblies for a large collection of genomes, and provides portals to the National Human Genome Research Institute's Encyclopedia of DNA Elements (ENCODE) and Neanderthal projects.

**VIDA** http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA3/VIDA.html

VIDA (Virus Database at University College London) contains a collection of homologous protein families derived from open reading frames from complete and partial virus genomes. Conserved sequence regions of potential functional importance are identified and can be retrieved as sequence alignments. Taxa included: Herpesviridae, Poxviridae, Papillomaviridae, Coronaviridae, and Arteriviridae.

**VIPERdb** http://viperdb.scripps.edu

VIPERdb is a database for icosahedral virus capsid structures operated by the Scripps Research Institute (US). The emphasis of the resource is on providing data from structural and energetic analyses on these systems, and high quality renderings for visual exploration. VIPERdb is a training/service and dissemination component of the National Institutes of Health - Multiscale Modeling Tools for Structural Biology program.

**Viral Bioinformatics Resource Center** http://athena.bioc.uvic.ca/ [formerly hosted at http://www.vbrc.org/]

The VBRC database currently operated by the University of Victoria contains databases of viral genomic information on large DNA viruses: Arenaviridae, Bunyaviridae, Flaviviridae, Filoviridae, Paramyxoviridae, Poxviridae, and Togaviridae including 30 genera, 209 species, 4958 strains, and 70732 genes. Domain-specific subsets of the VBRC are available for Poxviruses, Hepatitis C viruses, and Dengue viruses from legacy web sites whose future status remains uncertain: www.poxvirus.org, www.HCVdb.org, and www.DengueDb.org.

**ViralZone** http://expasy.org/viralzone/

Virus genomics resource and database supported by the Swiss Institute of Bioinformatics.

**VirGen** http://bioinfo.ernet.in/virgen/virgen.html

VirGen is a relational database that includes complete viral genome sequences, derived data, and data mining tools. VirGen is a relational database that includes complete viral genome sequences, derived data, and data mining tools. VirGen includes data for 25 viral families that include 2475 genomes and 22006 annotated proteins.

**Virulence Factors Database (VFDB)** http://www.mgc.ac.cn/VFs/

The VFDB contains database contains sequence data on 24 bacterial pathogens, virulence-associated genes and pathogenicity islands. The database is supported by [People's Republic of China] State Key Laboratory for Moleclular Virology and Genetic Engineering in Bejing, China.

**Virus Pathogen Resource (ViPR)** http://www.viprbrc.org/brc/home.do?decorator=vipr

The Virus Pathogen Database and Analysis Resource (ViPR) is an open-access resource supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health. ViPR provides genome data and metadata on 13 families of RNA and DNA viruses, with a variety of analytical and visualization tools. It is a virus-specific subcomponent of the NIAID Pathogen Portal


**Defense Threat Reduction Agency**
8725 John J. Kingman Road • Stop 6201
Fort Belvoir, Virginia 22060-6201
www.dtra.mil