Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Virology

The search for meaning in virus discovery

Peter Daszak[1] and W Ian Lipkin[2]

The rate of new virus discovery is increasing dramatically with improvements in sequencing and other molecular diagnostic platforms, and investments in sample collection and analysis. However, progress has been more limited in identification and implication of infectious agents that pose threats to human health and welfare. Here we review strategies for targeting research to enable efficient significant virus discovery.

Addresses
[1] EcoHealth Alliance, 460 West 34th Street, New York, USA
[2] Columbia University, 722 West 168th Street, 10032 New York, USA

Corresponding author: Lipkin, W Ian (wil2001@columbia.edu)

Current Opinion in Virology 2011, **1**:620–623

This review comes from a themed issue on
Emerging viruses
Edited by JS Malik Peiris and Colin Parrish

Available online 4th November 2011

1879-6257/$ – see front matter

DOI 10.1016/j.coviro.2011.10.010

You mean my whole fallacy is wrong?
Marshall McLuhan in Woody Allan's 'Annie Hall', 1977

Introduction

The emergence of new viral infections is a global threat to public health that is significant and increasing. Analysis of recent trends in emerging diseases shows that the number of new emerging infectious diseases (EIDs) has been growing decade-by-decade, and will continue to grow [1••]. This trend will be exacerbated by increased dependence on travel networks and globalized trade, the lack of investments in surveillance, diagnostics, therapeutics and vaccines. Emerging pathogens cause significant mortality (e.g. HIV/AIDS) as well as and high economic costs. For example, the emergence of SARS in 2001 may have cost between $30 and $50 billion, and estimates for the cost of an H5N1 pandemic are in the hundreds of billions of dollars [2].

Given limited global resources to tackle the threat of emerging pathogens, efforts to shift the focus of our global surveillance and diagnostic capacity to pre-empting pandemics are crucial for global health [3]. Investments that fuse modeling infectious disease emergence and pathogen discovery are based on the premise that focused sample collection joined with molecular surveillance technologies are the most effective strategy to identify novel emerging pathogens. Although these investments continue to bear fruit they also yield considerable chatter. The challenge is to refine our strategies to enhance the probability of the first. This review outlines our thoughts on how this can be achieved.

Modeling geographical EID 'hotspots' to target surveillance

Shifting the intervention from post-pandemic response to pre-pandemic prevention is a significant challenge. Most of the high impact emerging pathogens are viral zoonoses that have wildlife reservoir hosts [4,5]. Our approach, therefore, is to conduct sample collection and pathogen discovery in wildlife to attempt to identify likely zoonotic pathogens before they emerge in people. The global diversity of viruses that could spill over from wildlife is unknown and likely high, and the global distribution of this biodiversity is unknown. However, we can begin to work out its broad parameters. To get a crude estimate, we can consider that most emerging zoonoses have vertebrate non-human hosts (largely mammalian and avian). There are approximately 50,000 known vertebrate species, and if we estimate, conservatively, that each vertebrate species carries 20 endemic, as-yet uncharacterized viruses, then there is a global diversity of 1,000,000 viruses. With only approximately 2000 different species of viruses identified, we can crudely say that we underestimate the zoonotic pool by at least 99.8%. We can apply the same approach to different wildlife groups, for instance bats (Order Chiroptera), which are known to be the reservoirs for SARS-like corona-, Hendra, Nipah, Ebola, Marburg, and rabies viruses, and other emerging pathogens [6]. With around 1000 known bat species, the global diversity of unknown bat viruses is conservatively 20,000 species, and we have probably characterized less than 200.

To use our global pathogen discovery resources most effectively, we need to target the regions where (1) wildlife biodiversity is highest, and (2) where the factors that cause diseases to emerge occur. To do this, we constructed a database of all known emerging diseases, and identified all available data on the likely location and timing of the first spillover. We accounted for biases in reporting effort, which would otherwise skew the data. We were able to use this database to address some key questions in emerging disease biology [1••]. First, we demonstrated that the number of emerging infectious diseases is rising over time. Second, we showed that emerging zoonoses are rising at the fastest rate and, during the last decade analysed, represented the dominant group of EIDs.

We have used our database approach to test two important hypotheses: (1) That disease emergence is an 'anthropogenic' process (i.e. is caused by human changes to demography, the environment, and other factors; (2) That we can use this approach to create a predictive map of emerging disease 'hotspots'—the regions most likely to cause the next new emerging disease. To test these hypotheses, we used a simple approach to deal with the issue of not having accurate data on the geographic distribution of unknown pathogens – a crucial part of the equation of predicting the next disease. In our analyses, we assumed that each mammalian species harbors an equal number of host-specific pathogens. With this assumption, we were able to use the global distribution of wildlife diversity as a proxy for the global distribution of unknown zoonotic pathogen diversity. This approach allowed us to show definitively that disease emergence is correlated with human activity on the planet, that is it is a product of human environmental change, and demographic changes. It also allowed us to identify the geographic regions on the planet where these factors and wildlife biodiversity come together to create the right conditions for disease emergence. These EID 'hotspots' are primarily in the Tropics for zoonoses from wildlife, and include some regions of Europe and North America for other types of pathogens (e.g. food-borne infections).

Targeting key wildlife species and high-risk human populations

The hotspots maps provide a way to use limited global resources to target sample collection from the places on the planet where the next pandemic is most likely to originate. Within those regions, wildlife biodiversity is highest, and it is clearly not feasible to sample every species of vertebrate across these large swathes of the Tropics. To identify the wildlife species that harbor the greatest proportion of potential novel zoonoses, we need to consider two factors: phylogeny and contact. Logically, it is more likely that a virus harbored by a wildlife species phylogenetically closely related to humans will be able to replicate in human cells following exposure. However, testing this hypothesis has not been the focus of a great deal of research effort, other than studies in plants [7], bats [8], and insects [9]. With large databases of virus-host relationships, global datasets on wildlife host traits and viral phylogeny, extending these studies to produce predictive models should be plausible. Ultimately these approaches could provide predictions of the risk of a novel agent discovered in wildlife spilling over into people. Importantly, they will provide a way to target the wildlife species with the highest potential for discovering novel pathogens of zoonotic potential.

Viral spillover is also a product of the degree of contact that a wildlife host has with human populations. For example, pastoral communities on the edges of forests in Brazil, Borneo or Central Africa likely have a high degree of exposure to fomites from a high diversity of rodent species. If contact is more crucial than phylogeny, then viral discovery programs should target rodents rather than primates in hotspot countries. To test these ideas, and develop usable predictions, mathematical models that include global datasets on human contact with wildlife are required. Currently these do not exist. Our approach has been to use proxies for contact, based on the amount of anthropogenic disturbance within a region, in addition to the density of human population. These can then be correlated with results from large scale surveys of human populations across different gradients of disturbance to get more accurate data on likely wildlife exposure.

This research aims to provide better support for decision making for viral discovery and surveillance. In the absence of these analyses, decisions on which wildlife species to target for sample collection, and which human populations to conduct surveillance on are made based on assumptions about typical emerging diseases. For example, there is a great deal of interest in bush meat hunters because it is logical that their exposure to wildlife is intimate. A more targeted approach would be to estimate the intensity of exposure, the frequency of exposure and the species that different groups make contact with in EID hotspots. For example, if a hunter only catches one primate a week, is that person at higher risk of viral spillover than someone who lives on the forest edge and gets regular exposure to primate or rodent fomites and feces? Efforts to effectively target the first case clusters of a newly emerging pathogen could involve surveillance of other high-risk communities such as forest workers, road builders, miners, livestock production workers, abattoir workers, field hands and others. However, analyses of the relative risk of exposure in each of these populations are crucial to better inform surveillance, and give the most effective use of resources. These analyses might take the form of field studies to measure human-wildlife contact along land use gradients in hotspot regions.

Molecular surveillance

The advent of culture-independent tools for microbial diagnostics, surveillance and discovery has revolutionized medicine and biology. Some applications have clearly had enormous utility, for example monitoring drug responses to antiviral therapy in HIV/AIDS or hepatitis C, or rapid characterization of agents associated with outbreak of infectious disease like West Nile virus, SARS coronavirus or highly pathogenic *E. coli*. However, others have simply flooded databases with sequences of known and novel microflora. Although one can argue that all sequence data have value, in an era of diminishing resources, focused investment is increasingly important.

A staged strategy for pathogen surveillance and discovery

The most efficient and successful pathogen searches typically begin with samples collected from geographically and temporally clustered individuals with acute disease. Key advantages in cluster analyses are that one can test for a statistically significant association with disease, and one has many samples to examine. Hence, if the load is too low for detection in some samples using discovery methods, one can return to those samples with specific sensitive methods after finding a candidate in sample with a higher viral load. An additional advantage is that there may be an opportunity to collect sera at later time points to enable assays for adaptive immune responses indicative of current infection. The most important individual(s) in the process are the clinicians and epidemiologists who appreciate the anomaly of the disease cluster and collect the samples distributed to research laboratories. We and others have been successful in identifying new pathogens in single cases of disease; however, the bar for proof of causation is higher. The agent should be present at high concentrations in the affected tissue, seroconversion should be demonstrated, and it is helpful if there is precedent for a similar agent causing a similar disease in the same host or another. Confidence in a causal relationship between a candidate pathogen and a disease is enhanced by fulfillment of Kochs' Postulates (i.e. demonstration of the presence of an agent in all cases of a disease and not in the absence of disease, replication of disease following ex vivo cultivation and introduction into a naïve host); however, this is not always feasible. Other factors may confound recognition of a link between a microbe and disease. Individuals may vary in susceptibility because of genetic factors (e.g. absence of the CCR5 receptor confers resistance to HIV infection), age (e.g., the elderly are more prone to West Nile virus encephalitis), or previous exposures that may either prevent or enhance disease (e.g., vaccinia virus infection results in protective immunity to *Variola*; exposure to one Dengue serotype may increase risk of hemorrhagic fever on exposure to another serotype). Immune or toxin mediated effects can occur distal to the site of infection. For a recent review of principles employed in pathogen discovery and proving causal relationships the reader may wish to consult a recent review on Microbe Hunting [10••].

In instances where the focus is on surveillance for specific pathogens in natural reservoirs, vectors or at-risk human or animal populations singleplex molecular or serological assays are sufficient. Such assays can be both sensitive and inexpensive. However, as the costs drop for MassTag PCR, microarrays and high throughput sequencing, multiplex assays are increasingly used as primary tools for syndromic surveillance, studies of microbial diversity and discovery. Sensitivity can be a challenge with multiplex platforms. In multiplex PCR reactions sensitivity decreases with increasing primer complexity. Microarrays and deep sequencing typically employ unbiased amplification methods wherein host and microbial sequences can compete for polymerase and nucleotides. Improvements in sensitivity can be achieved by using methods that deplete host DNA, ribosomal and mitochondrial sequences through enzymatic digestion or subtractive hybridization before unbiased amplification. One can also pursue positive selection using oligonucleotides representing microbes of interest. With high throughput sequencing bioinformatics is frequently the weakest link. The least expensive platforms (e.g. Illumina) typically produce shorter sequence reads that can be difficult to assemble. More expensive platforms (e.g. 454 Life Sciences) typically yield longer but fewer reads; hence, assembly is easier but at a cost of lower coverage. One solution is to join such platforms and use the longer reads to establish a scaffold that can be completed using the shorter reads. Single molecule sequencers (e.g. Pacific Biosciences) are in development that may provide the best of both worlds, low cost and long, contiguous strings of sequence.

Sequence analysis

Acquisition of sequence is only the first step in determining the identity of an agent, its provenance and relationship to other microbes. Most initial screens are performed using alignment programs that test for similarity at the nucleotide and amino acid levels. However, nucleotide composition and motif based programs may succeed where alignments fail. In discovery of the piscine reovirus of salmon for example, assays of nucleotide composition and order enabled detection of two gene segments not found by alignment. Nucleotide composition has also been used to resolve determine whether sequences found in the gastrointestinal tract of mammals represent viruses infecting the host, animal or vegetal matter consumed by the host, or an inadvertent insect passenger contaminating food stuffs. Unlike bacteria where pathogenicity islands are readily defined viruses typically don't have sequences with obvious functional correlates. Indeed, single nucleotide changes may have a profound impact on pathogenicity. Whether genetic signatures can be discovered that predict pathogenicity or cross species transmission remains to be seen.

Acknowledgements

Peter Daszak's work is supported by NIAID Non-biodefense emerging infectious disease research opportunities award 1 R01 AI079231, an NIH/NSF 'Ecology of Infectious Diseases' award from the Fogarty International Center 2R01-TW005869, the Rockefeller Foundation, Google.org, NSF Human and Social Dynamics 'Agents of Change' award (SES-HSD-AOC BCS–0826779), and generous support of the American people through the United States Agency for International Development (USAID) Emerging Pandemic Threats PREDICT. The contents are the responsibility of the authors and do not necessarily reflect the views of USAID or the United States Government. Dr. Lipkin's work is supported by grants from the National Institutes of Health (AI057158, AI0793231, AI070411, EY017404), Bill and Melinda Gates Foundation, USAID PREDICT, and Defense Threat Reductions Agency.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. •• Jones KE, Patel N, Levy M, Storeygard A, Balk D, Gittleman JL, Daszak P: **Global trends in emerging infectious diseases**. *Nature* 2008, **451**:990-994.

This paper shows that analysis of the underlying causes of disease emergence can be used to predict where on the planet the next emerging disease is most likely to originate.

2. The World Bank: **Avian and human pandemic influenza – economic and social impacts**. *Press Release* (2005).

3. **Emerging pandemic threats: Program overview**: (2010). http://www.usaid.gov/our_work/global_health/home/News/ai_docs/emerging_threats.pdf.

4. Lloyd-Smith JO, George D, Pepin KM, Pitzer VE, Pulliam JRC, Dobson AP, Hudson PJ, Grenfell BT: **Epidemic dynamics at the human-animal interface**. *Science* 2009, **326**:1362-1367.

5. Keusch GT, Pappaioanou M, Gonzalez MC, Scott KA, Tsai P (Eds): *Sustaining global surveillance and response to emerging zoonotic diseases.*. Washington, D.C.: The National Academies Press; 2009.

6. Calisher CH, Childs JE, Field HE, Holmes KV, Schountz T: **Bats: Important reservoir hosts of emerging viruses**. *Clin Microbiol Rev* 2006, **19**:531-545.

7. Gilbert GS, Webb CO: **Phylogenetic signal in plant pathogen-host range**. *Proc Natl Acad Sci U S A* 2007, **104**:4979-4983.

8. Streicker DG, Turmelle AS, Vonhof MJ, Kuzmin IV, McCracken GF, Rupprecht CE: **Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats**. *Science* 2010, **329**:676-679.

9. Longdon B, Hadfield JD, Websterm CL, Obbard DJ, Jiggins FM: **Host phylogeny determines viral persistence and replication in novel hosts**. *PLoS Pathogens* 2011, **7**:e1002260.

10. •• Lipkin WI: **Microbe hunting in the 21st century**. *Proc Natl Acad Sci U S A* 2009, **106**:6-7.

This paper gives further details on the use of a staged pathogen discovery platform for novel microbes.