OPEN ACCESS Freely available online

PLoS BIOLOGY

Perspective

Can an Infectious Disease Genomics Project Predict and Prevent the Next Pandemic?

Rajesh Gupta¤*, Mark H. Michalski¤, Frank R. Rijsberman

Google.org, Mountain View, California, United States of America

We believe that there is great potential in the systematic application of genomics, proteomics, and bioinformatics to infectious diseases, and that this potential has yet to be fully realized. We suggest that the international community unite under an Infectious Disease Genomics Project, analogous to the Human Genome Project, with a goal of a comprehensive, open-access system of genomic information to accelerate scientific understanding and product development in the very settings where diseases have the highest probability of emerging. If properly structured, such an approach could shift fundamentally the global response to emerging infectious diseases.

Genomics Is Systematically Transforming Medicine

The "Genomic Revolution" has transformed our vision and understanding of how living organisms and systems interact with each other and with the environment [1]. Increasingly, the science of genomics serves as the foundation for translational research for advancing the management of many important diseases [2–7]. Decreasing costs and increasing throughput of new technologies has made possible multinational collaboration on large-scale projects such as the Human Microbiome Project and the 1000 Genomes Project [8–10]. Infectious disease management is also transforming thanks to molecular technologies as seen in HIV [11,12], tuberculosis [13,14], malaria [15,16], and other neglected tropical diseases [17,18]. Discovering novel pathogens and elucidating the implications of genetic variation among existing pathogens [19,20] is critical for rapidly mitigating pandemic threats, as demonstrated recently with severe acute respiratory syndrome (SARS) [21,22] and avian (H5N1) and pandemic H1N1 2009 influenza (commonly referred to as "swine flu") [23–26].

To fully harness the benefit of genomics in infectious diseases, a chain of overarching activities must occur. First, understanding the dynamics of infectious diseases through the genomics lens requires a tremendous amount of integrated comparative sequence, expression, epigenetic, and proteomic data from a variety of pathogens (bacteria, virus, protozoa, fungi), vectors (arthropod and avian sources), reservoirs (non-human mammals, environment) and human hosts. Second, generating, collating, organizing, and curating these data is an essential public health task. Third, translating this information to tools to improve surveillance and response mechanisms is critical to effectively impact disease management.

If this bench-to-beside chain of activities were optimized, we envision that the following could occur:

- Fully annotated genomes of all known pathogens, vectors, non-human hosts, and reservoir species, as well as a large number of candidate microbes in families that have a high risk of generating future pathogens, are held in public open-access databases such as GenBank.
- A "Genomic search" of all available contextual information, from sample origins through to published analyses, is as simple as a Google search.
- Sequencing and other molecular technologies are everyday tools-of-the-trade in every district hospital and laboratory in hotspots of emerging infectious disease, such as southeast Asia and sub-Saharan Africa.
- Automated molecular diagnostic assays are low-cost, reduced at least to the size of a smart mobile phone, and can return definitive diagnoses of a range of specialized known pathogen panels at the point of care.
- A range of products that use infectious disease genomic information routinely—such as vector maps, early warning systems, diagnostics, vaccines, and drugs—contribute to the prediction and prevention of epidemics.

While progress is occurring in each of these areas, the outputs—which are needed today—are far from complete.

Creating an Infectious Disease Genomics Project (IDGP)

We believe that accelerated advances in the area of infectious diseases can occur under a global collaborative framework composed of discrete and delineated activities between the public and private sectors among resource-wealthy and resource-limited settings. The Human Genome Project (HGP) was a pioneering international effort that helped unlock the power of genomics for human health

The Perspective section provides experts with a forum to comment on topical or controversial issues of broad interest.

Citation: Gupta R, Michalski MH, Rijsberman FR (2009) Can an Infectious Disease Genomics Project Predict and Prevent the Next Pandemic? PLoS Biol 7(10): e1000219. doi:10.1371/journal.pbio.1000219

Published October 26, 2009

Funding: Google.org is financially supported through its parent company, Google.com. At the time this manuscript was developed, RG was an employee of Google.org and MM was a consultant to Google.org. The funder had no role in the decision to publish or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rgupta1@stanford.edu

¤ **Current address:** Stanford University, Stanford, California, United States of America

This article is part of the "Genomics of Emerging Infectious Disease" PLoS Journal collection (http://ploscollections.org/emerginginfectiousdisease/).

Author Summary

The world of genomics is transforming medicine, and is likely to influence the future development of new drugs, diagnostics, and vaccines. To date, the greater focus of genomics and medicine has been on conditions affecting resource-wealthy settings, primarily involving scientists and companies in those settings. However, we believe that it is possible to expand genomics into a more global technology that can also focus on diseases of resource-limited settings. This goal can be achieved if genomics is made a global priority. We feel one way to move in this direction is through a comprehensive approach to infectious diseases—i.e., an Infectious Disease Genomics Project—that would mirror the Human Genome Project. Without an active, unified effort specifically focused on allowing actors at any level to participate in the genomics revolution, infectious diseases that primarily affect the poor will likely not achieve the same level of scientific advancement as diseases affecting the wealthy.

[27,28]. This effort generated important information in part by having clear, targeted outcomes and by implementing a standard methodology across all participants. The HGP was a great impetus for progress seen thus far in genomics and health. Moreover, the HGP recognized that sequencing was just the first step in a much bigger process [26]. A similar effort for infectious diseases could, in our view, help predict and prevent the next pandemic.

To capitalize on existing successful efforts in the area of genomics and infectious diseases such as those by the Broad Institute, Genomics Standards Consortium, J Craig Venter Institute, the National Institute of Allergy and Infectious Diseases, and the Wellcome Trust Sanger Institute (to name a few), we urge the international community to unite its numerous activities under an Infectious Diseases Genomic Project (IDGP)—a coordinated, large-scale, international effort focused on the genomes of pathogens, vectors, hosts, and reservoirs and linked to end-point surveillance and response systems. Such a project could coordinate activities in four specific areas: generating data, linking data, analyzing data, and applying data (Figure 1).

Generating Data

At the outset, the IDGP would need to determine what the world requires in terms of genomic information. A standard approach to generating depth and diversity in genomic data is essential; beyond this, continuous real-time surveillance and characterization of evolving pathogens can help effectively forestall future epidemics/pandemics. Frontline work by consortiums, genome research centers, and individual laboratories has yielded baseline approaches in this area and a wealth of critical genomic information for many important infectious agents [29–34]. While each actor in the genomics field brings its own priority for targeting particular pathogens or diseases, a clear roadmap to generating a complete genomic picture of *all* infectious agents, emerging threats, hosts, and reservoirs, incorporating a broad range of investigators with varied technological capacity, would enhance both data generation and application. Such a process allows for community-level priority setting, thereby enabling smaller-scale laboratories to tailor projects to fit the needs of local communities while contributing to global efforts.

Linking Data

The data collected must be connected to all relevant information and analytical tools in a single, easy-to-use, open-source, real-time interface. Such a system would improve on current systems by: gathering data across the public domain and working with companies/institutions to harness information in the private domain; linking accurate, annotated sequencing information to functional genomic and proteomic/functional proteomic information; attaching scientific literature associated with all levels of information; and including a self-sustaining financial mechanism potentially based on royalties from commercial products generated from the use of this system.

Analyzing Data

The data need to be linked via large-scale, dynamic databases held in virtual servers allowing for collaboration and sharing while maintaining originating information for data rights and sovereignty. Concurrently, these data should be associated with a centralized collection of open-source bioinformatics tools capable of real-time operation in low- and high-speed computers and varying levels of internet connectivity. A single interface also would bring various sample collections together in formally structured biobanks that capture geospatial and context data to allow efficient scientific collaboration to take place. Centralizing the entire spectrum of information and analytic tools also allows researchers in resource-limited settings to participate in the genomics revolution without prohibitively costly machines, laboratories, and sample accessibility. Although we fully acknowledge

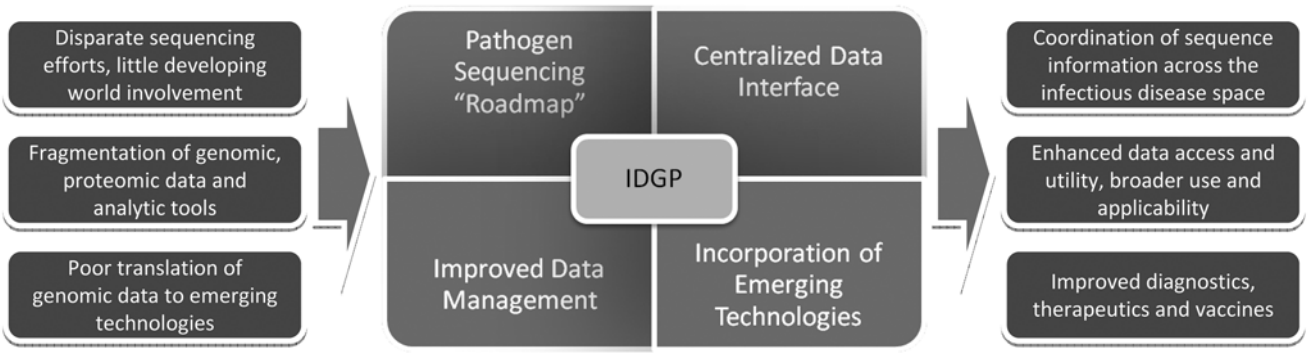

Figure 1. A coordinated Infectious Disease Genome Project (IDGP) could unify sequencing efforts, enhance data usability, and lead to essential tools for infectious disease management.
doi:10.1371/journal.pbio.1000219.g001

that internet connectivity is a requirement that is not currently available to all, rapid technical innovation and investment from cheap netbook computers to new fiber optic cables in Africa are changing that equation. This system could be facilitated by virtual community collaboration or crowd-sourcing, taking full advantage of networking tools such as Wikipedia, Facebook, Twitter, FusionTables, and PLoS.

Applying Data

Technological advances for basic scientific discovery (such as next-generation sequencers, microarrays, mass spectrometers, cell-based assay methods, and other tools for transcriptome, metabolome, and proteome discovery), novel techniques to increase throughput and/or decrease the cost of analysis, and applied clinical decision-making and surveillance tools (point-of-care diagnostics, rapid multi-pathogen assays) are in progress and should be supported actively. The IDGP should be informed by and incorporate emerging technology platforms to rapidly develop more accurate field diagnostics and to identify new opportunities for vaccine and drug development.

Moving beyond Discourse into Action

An IDGP is attainable if others share this vision, show leadership, and see the added value resulting from a coordinated effort. The HGP certainly was a more targeted effort and we acknowledge that an IDGP will have additional obstacles to overcome. Scientific disagreement over targets is bound to occur. Complications resulting from the proposed level of data sharing should not be underestimated, and care must be taken to ensure proprietary rights and acknowledgement when warranted. Adapting molecular genetic technologies to resource-limited settings is a significant challenge, but is occurring with some success. Bringing together a community of scientists and donors, each with their own objectives and goals, to work under a single framework, is a difficult proposition. Finally, there will be many who will find this perspective simply too grandiose. Leaps of progress also require big visions, however, and it may just be possible that the 2009 H1N1 influenza pandemic is a enough of a reminder of what is at stake to provide a catalyst for action.

Google.org has supported global public health through its "Predict and Prevent" initiative with the aim of using the power of information and technology to address emerging infectious diseases by helping the world to know where to look for these diseases, find the threats earlier, and respond to them faster [35]. Google.org has focused its support on sequencing and pathogen discovery activities, bringing genomic technologies to resource-limited settings in East Africa, improving surveillance networks and systems, and exploring how our core competence in internet search can assist the infectious diseases community [36].

As firm supporters of the open access model for scientific publication [37], Google.org is pleased to support this series of essays, The Genomics of Emerging Infectious Disease, in partnership with the Public Library of Science (PLoS) journals (*PLoS Biology*, *PLoS Computational Biology*, *PLoS Genetics*, *PLoS Medicine*, *PLoS Neglected Tropical Diseases*, and *PLoS Pathogens*), not only to help define the current state of the art in pathogen genomics, but also, we hope, to stimulate debate on priorities for research and technology development.

References

1. Yudell M, DeSalle R (2002) The genomic revolution: Unveiling the unity of life. Washington (D. C.): Joseph Henry Press. 272 p.
2. Langston AA, Malone KE, Thompson JD, Daling JR, Ostrander EA (1996) BRCA1 mutations in a population-based sample of young women with breast cancer. N Engl J Med 334: 137–142.
3. Futreal P, Liu Q, Shattuck-Eidens D, Cochran C, Harshman K, et al. (1994) BRCA1 mutations in primary breast and ovarian carcinomas. Science 266: 120–122.
4. Helgadottir A, Manolescu A, Thorleifsson G, Gretarsdottir S, Jonsdottir H, et al. (2004) The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. Nature Genetics 36: 233–239.
5. Wellcome Trust C (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
6. Consortium G (2007) New models of collaboration in genome-wide association studies: The Genetic Association Information Network. Nat Genet 39: 1045–1051.
7. Vigneri P, Wang J (2001) Induction of apoptosis in chronic myelogenous leukemia cells through nuclear entrapment of BCR-ABL tyrosine kinase. Nat Med 7: 228–234.
8. Gresham D, Kruglyak L (2008) Rise of the machines. PLoS Genet 4: e1000134. doi:10.1371/journal.pgen.1000134.
9. Spencer G (2008) Researchers establish international human microbiome consortium. NIH News. Available: http://www.nih.gov/news/health/oct2008/nhgri-16.htm. Accessed 19 September 2009.
10. Spencer G (2008) International consortium announces the 1000 Genomes Project. NIH News. Available: http://www.nih.gov/news/health/jan2008/nhgri-22.htm. Accessed 19 September 2009.
11. Martinez-Cajas JL, Wainberg MA (2008) Antiretroviral therapy: Optimal sequencing of therapy to avoid resistance. Drugs 68: 43–72.
12. Wilkinson KA, Gorelick RJ, Vasa SM, Guex N, Rein A, et al. (2008) High-throughput SHAPE analysis reveals structures in HIV-1 Genomic RNA strongly conserved across distinct biological states. PLoS Biol 6: e96. doi:10.1371/journal.pbio.0060096.
13. Smith CV, Sacchettini JC (2003) *Mycobacterium tuberculosis*: A model system for structural genomics. Curr Opin Struct Biol 13: 658–664.
14. Cockle PJ, Gordon SV, Lalvani A, Buddle BM, Hewinson RG, et al. (2002) Identification of novel *Mycobacterium tuberculosis* antigens with potential as diagnostic reagents or subunit vaccine candidates by comparative genomics. Infect Immun 70: 6996–7003.
15. Gonzales JM, Patel JJ, Ponmee N, Jiang L, Tan A, et al. (2008) Regulatory hotspots in the malaria parasite genome dictate transcriptional variation. PLoS Biol 6: e238. doi:10.1371/journal.pbio.0060238.
16. Ekland EH, Fidock DA (2007) Advances in understanding the genetic basis of antimalarial drug resistance. Curr Opin Microbiol 10: 363–370.
17. Beaty BJ, Prager DJ, James AA, Jacobs-Lorena M, Miller LH, et al. (2009) From Tucson to genomics and transgenics: The Vector Biology Network and the emergence of modern vector biology. PLoS Negl Trop Dis 3: e343. doi:10.1371/journal.pntd.0000343.
18. Hertz-Fowler C, Figueiredo LM, Quail MA, Becker M, Jackson A, et al. (2008) Telomeric expression sites are highly conserved in *Trypanosoma brucei*. PLoS ONE 3: e3527. doi:10.1371/journal.pone.0003527.
19. Wolfe N, Heneine W, Carr J, Garcia A, Shanmugam V, et al. (2005) Emergence of unique primate T-lymphotropic viruses among central African bushmeat hunters. Proc Natl Acad Sci U S A 102: 7994–7999.
20. Palacios G, Druce J, Du L, Tran T, Birch C, et al. (2008) A new arenavirus in a cluster of fatal transplant-associated diseases. N Engl J Med 358: 991–998.
21. Grant P, Garson J, Tedder R, Chan P, Tam J, et al. (2003) Detection of SARS coronavirus in plasma by real-time RT-PCR. N Engl J Med 349: 2468.
22. Marra M, Jones S, Astell C, Holt R, Brooks-Wilson A, et al. (2003) The genome sequence of the SARS-associated coronavirus. Science 300: 1399–1404.
23. Gu J, Xie Z, Gao Z, Liu J, Korteweg C, et al. (2007) H5N1 infection of the respiratory tract and beyond: A molecular pathology study. Lancet 370: 1137–1145.
24. Zhao Z-M, Shortridge KF, Garcia M, Guan Y, Wan X-F (2008) Genotypic diversity of H5N1 highly pathogenic avian influenza viruses. J Gen Virol 89: 2182–2193.
25. Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, et al. (2009) Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science 325: 197–201.
26. Shinde V, Bridges CB, Uyeki TM, Shu B, Balish A, et al. (2009) Triple-reassortant swine influenza A (H1) in humans in the United States, 2005–2009. N Engl J Med 360: 2616–2625.
27. Consortium IHGS (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
28. Collins FS, Morgan M, Patrinos A (2003) The Human Genome Project: Lessons from large-scale biology. Science 300: 286–290.
29. Wellcome Trust Sanger Institute (2009) Pathogen genomics [Web site]. Available: http://www.sanger.ac.uk/Projects/Pathogens/. Accessed 11 August 2009.

30. National Institute of Allergy and Infectious Disease (2009) Microbial Genome Sequencing Centers: Completed NIAID-Supported Sequencing Projects. Available: http://www3.niaid.nih.gov/research/resources/mscs/completed.htm. Accessed 11 August 2009.
31. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393: 537–544.
32. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419: 498–511.
33. Greene JM, Collins F, Lefkowitz EJ, Roos D, Scheuermann RH, et al. (2007) National Institute of Allergy and Infectious Diseases bioinformatics resource centers: New assets for pathogen informatics. Infect Immun 75: 3212–3219.
34. Field D, Garrity G, Gray T, Morrison N, Selengut J (2008) The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 26: 541–547.
35. Google.org (2008) Predict and Prevent initiative. Available: http://www.google.org/predict.html. Accessed 19 September 2009.
36. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. (2009) Detecting influenza epidemics using search engine query data. Nature 457: 1012–1014.
37. Gass A (2004) Open access as public policy. PLoS Biol 2: e353. doi:10.1371/journal.pbio.0020353.