

COMMENT

BIG DATA Three books on exabytes, in academia, business and governance **p.480**

DESERTS On the natural and cultural abundance of arid places **p.482**

REPRODUCIBILITY Validation is still vital, even if antibodies are recombinant **p.483**



OBITUARY Robert A. Berner, carbon-cycle geochemist, remembered **p.484**

SAMUEL ARANDA/PANOS



Nurses at the Kenema hospital in Sierra Leone, which contributed data to early efforts to sequence the genome of the Ebola virus from the West Africa outbreak.

Make outbreak research open access

Establish principles for rapid and responsible data sharing in epidemics, urge **Nathan L. Yozwiak, Stephen F. Schaffner and Pardis C. Sabeti.**

Last April, five months into the largest Ebola outbreak in history, an international group of researchers sequenced three viral genomes, sampled from patients in Guinea¹. The data were made public that same month. Two months later, our group at the Broad Institute in Cambridge, Massachusetts,

sequenced 99 more Ebola genomes, from patients at the Kenema Government Hospital in Sierra Leone.

We immediately uploaded the data to the public database GenBank (see go.nature.com/aotpbk). Our priority was to help curb the outbreak. Colleagues who had worked with us for a decade were at the

front lines and in immediate danger; some later died. We were amazed by the surge of collaboration that followed. Numerous experts from diverse disciplines, including drug and vaccine developers, contacted us. We also formed unexpected alliances — for instance, with a leading evolutionary virologist, who helped us to investigate ►

► when the strain of virus causing the current outbreak arose.

The genomic data confirmed that the virus had spread from Guinea to Sierra Leone, and indicated that the outbreak was being sustained by human-to-human transmission, not contact with bats or some other carrier. They also suggested new probable routes of infection and, importantly, revealed where and how fast mutations were occurring². This information is crucial to designing effective diagnostics, vaccines and antibody-based therapies.

What followed was three months of stasis, during which no new virus sequence information was made public (see 'Gaps in the data'). Some genomes are known to have been generated during this time from patients treated in the United States³. The number is likely to have been much larger: thousands of samples were transferred to researchers' freezers across the world.

In an increasingly connected world, rapid sequencing, combined with new ways to collect clinical and epidemiological data, could transform our response to outbreaks. But the power of these potentially massive data sets to combat epidemics will be realized only if the data are shared as widely and as quickly as possible. Currently, no good guidelines exist to ensure that this happens.

SPEED IS EVERYTHING

Researchers working on outbreaks — from Ebola to West Nile virus — must agree on standards and practices that promote and reward cooperation. If these protocols are endorsed internationally, the global research community will be able to share crucial

information immediately wherever and whenever an outbreak occurs.

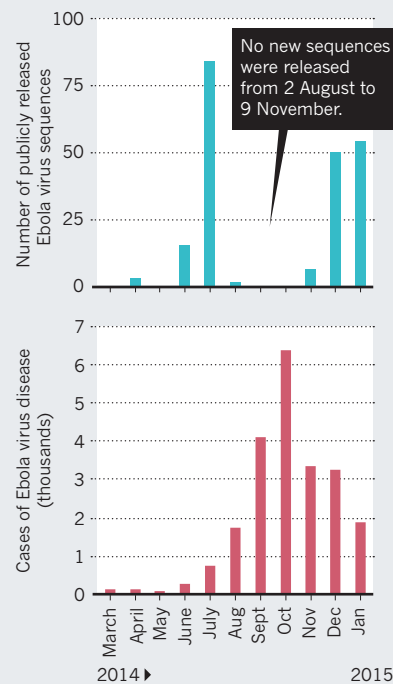
The rapid dissemination of results during outbreaks is sporadic at best. In the case of influenza, an international consortium of researchers called GISAID established a framework for good practice in 2006. Largely thanks to this, during the 2009 H1N1 influenza outbreak, the US National Center for Biotechnology Information created a public repository that became a go-to place for the community to deposit and locate H1N1 sequence information⁴. By contrast, the publishing of sequence information in the early stages of the 2012 Middle East respiratory syndrome (MERS) outbreak in Saudi Arabia highlighted uncertainties about intellectual-property rights, and the resulting disputes hampered subsequent access to samples.

Sharing data is especially important and especially difficult during an outbreak. Researchers are racing against the clock. Every outbreak can mobilize a different mixture of people — depending on the microbe and location involved — bringing together communities with different norms, in wildly different places. Uncertainties over whether the information belongs to local governments or data collectors present further barriers to sharing. So, too, does the absence of patient consent, common for data collected in emergencies — especially given the vulnerability of patients and their families to stigmatization and exploitation during outbreaks. Ebola survivors, for instance, risk being shunned because of fears that they will infect others.

Fortunately, useful models for responsible data sharing have been developed

GAPS IN THE DATA

Genome sequences from the West Africa outbreak of Ebola virus were first made publicly available in April 2014. Since 99 genomes were released in July, data sets have been shared sporadically, even though more are known to have been generated.



by the broader genomics community. In 1996, at a summit held in Bermuda, the heads of the major labs involved in the Human Genome Project agreed to submit DNA sequence assemblies of 1,000 bases or more to GenBank within 24 hours of producing them^{5,6}. In exchange, the sequencing centres retained the right to be the first to publish findings based on their own complete data sets, by laying out their plans for analyses in 'marker' papers.

This rapid release of genomic data served the field well. New information on 30 disease genes, for instance, was published before the release of the complete human genome sequence. Since 1996, the Bermuda principles have been extended to other types of sequence data and to other fields that generate large data sets, such as metabolite research.

GUIDELINES FOR SHARING

More-recent policies on data release similarly seek to align the interests of different parties, including funding agencies, data producers, data users and analysts, and scientific publishers. Since January, for example, the US National Institutes of Health has required grantees to make large-scale genomics data public by the time of publication at the latest, with earlier deadlines for some kinds of data⁷.



Pilgrims in Saudi Arabia try to protect themselves from Middle East respiratory syndrome (MERS) virus.

We urge those at the forefront of outbreak research to forge similar agreements, taking into account the unique circumstances of an outbreak.

First, incentives and safeguards should be created to encourage people to release their data quickly into the public domain. One possibility is to request that data users (and publishers) honour the publication intentions of data producers — the questions and analyses that they want to pursue themselves — for, say, six months. These intentions could be broadcast through several channels, including citable marker papers, disclaimer notices on data repositories such as GenBank, and online forums, such as virological.org and the EpiFlu database. Alternatively, data producers could publish an announcement about their data and their intentions on online forums as a resource that can be used by others as long as they cite the original source.

Second, ethical, rigorous and standardized protocols for the collection of samples and data from patients should be established to facilitate the generation and sharing of that information. A global consortium involving the leading health and research agencies and the ministries of health of engaged nations should work together towards establishing these. Ethicists should be involved to safeguard subjects' privacy and dignity. Biosecurity experts will also be needed to address potential dual-use research and other safety concerns. A helpful analogue is the approach used by the Human Heredity and Health in Africa (H3Africa) Initiative, which aims to apply genomics to improving the health of African populations. Since August 2013, H3Africa has used standard consent-form guidelines⁸ for collecting DNA samples from subjects for genomic studies, regardless of their country of origin.

Lastly, any preparation for future outbreaks should include provisions for rapidly building new bridges and establishing community norms. Successful collaborations in genomics and historical data-sharing agreements have tended to involve a fairly stable group of individuals and organizations, making norms of behaviour relatively easy to establish and sustain. By contrast, outbreaks can involve a new cast of characters each time, and cases in which the pathogen is new to science call for whole new fields of research.

THE KENEMA WAY

As a first step, we call on health agencies such as the World Health Organization, the US Centers for Disease Control and Prevention and Médecins Sans Frontières, as well as genome-sequencing centres and other research institutions, to convene a meeting this year — similar to that held in Bermuda



Quarantine officers rush to test passengers at Tokyo's Narita airport amid the 2009 swine-flu outbreak.

in 1996. Attendees must include scientists, funders, ethicists, biosecurity experts, social scientists and journal editors.

We urge researchers working on outbreaks to embrace a culture of openness. For our part, we have released all our sequence data as soon as it has been generated, including that from several hundred more Ebola samples we recently received from Kenema. We have listed the research questions that we are pursuing at virological.org and through GenBank, and we plan to present our results at virological.org as we generate them, for others to weigh in on. We invite people either to join our publication, or to prepare their own while openly laying out their intentions online. We have also made clinical data for 100 patients publicly available and have incorporated these into a user-friendly data-visualization tool, Mirador, to allow others to explore the data and uncover new insights.

“We urge researchers working on outbreaks to embrace a culture of openness.”

Kenema means ‘translucent, clear like a river stream’ or ‘open to the public gaze’⁹. To honour the memory of our colleagues who died at the forefront of the Ebola outbreak, and to ensure that no future epidemic is as devastating, let’s work openly in outbreaks. ■

Nathan L. Yozwiak and Stephen F. Schaffner are senior staff scientists, and **Pardis C. Sabeti** is associate professor, at the Broad Institute and Harvard University in Cambridge, Massachusetts, USA. e-mail: nyozwiak@broadinstitute.org

1. Baize, S. *et al.* *N. Engl. J. Med.* **371**, 1418–1425 (2014).
2. Gire, S. K. *et al.* *Science* **345**, 1369–1372 (2014).
3. Vogel, G. *Science* **346**, 684–685 (2014).
4. Bogner, P. *et al.* *Nature* **442**, 981 (2006).
5. Marshall, E. *Science* **291**, 1192 (2001).
6. Toronto International Data Release Workshop. *Nature* **461**, 168–170 (2009).
7. US National Institutes of Health. ‘NIH Genomic Data Sharing Policy’ (27 August 2014); available at <http://go.nature.com/i88ot5>
8. H3Africa. *H3Africa Guidelines for Informed Consent* (H3Africa, 2013).
9. Jackson, M. D. *Life Within Limits: Well-being in a World of Want* (Duke Univ. Press, 2011).