Applied & Translational Genomics 3 (2014) 109–110

Contents lists available at ScienceDirect

Applied & Translational Genomics

journal homepage: www.elsevier.com/locate/atg

Data acquisition and data/knowledge sharing in global genomic studies ☆

Charles Rotimi [a,*], Nicola Mulder [b,*]

[a] Center for Research on Genomics and Global Health, National Human Genome Research Institute, NIH, Bethesda, MD, USA
[b] Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, South Africa

Multi-center and multi-national collaborations are generating genomic, clinical and epidemiological datasets on ever increasing scales across global populations. Early examples include the international HapMap project and ongoing efforts include the 1000 Genomes Project, the genotyping and sequencing of whole populations or significant numbers of national of certain countries. The new initiatives include the UK10K project (http://www.uk10k.org), which aims to sequence thousands of genomes in Britain to identify rare variants and produce a sequence variation resource for future studies, the Qatar Genome Project, which will map the genomes of a large group of Qataris to chart a road map for future treatment through personalized medicine, and the Estonian Genome Project, which aims to generate data from 70% of Estonia's population. As a result of the broad sharing and analysis of these datasets, we are gaining new knowledge that is shedding novel insights into human history and health, such as the patho-biology of some of the world's most intractable diseases, and strategies for improving the efficacy and safety of treatments for several global diseases. Open sharing of data (with public access either directly or via data access committee, depending on the consent conditions) is thus extremely important for fully exploiting results and advancing research.

However, to maximize the benefits for the world populations in terms of improved treatments for genetic disorders that access to genomic data can facilitate, while minimizing the risks of broad genomic data sharing, analysis and interpretation, it is important to recognize some challenges. The major impediments to global data sharing are protection of participants and technical challenges in accessing data from disparate sources. Here we highlight three issues; first, because human genomic, clinical and epidemiological data can contain very sensitive information at the individual and population levels, there is a need for both rigorous institutional ethical review and the implementation of an informed consent process that protects participants, but at the same time provides flexibility to use the data for broader research. To ensure continued participants' protection, data is usually anonymized. However, if accompanied by enough demographic or phenotypic data, anonymization is not always 100% effective. In addition, many projects that have been undertaken did not include broad consent for data sharing and access. In some of these cases selected aspects of the data may still be shareable, but only if it is made publicly accessible via some means.

In most cases, data sits behind data access committees and once access is granted, raw or processed files are made available. However, some computational experience is usually required to be able to use the data, namely to mine it or perform meta-analysis. Further, in most cases technological infrastructure is needed to permit access to specific segments of a dataset. The Global Alliance for Genomics and Health (G4GH: http://genomicsandhealth.org/) initiative is developing Application Programming Interfaces (APIs) that will address these types of requirements and in turn facilitate data sharing across federated resources. These APIs will, for example, enable researchers to compare variants from different studies or simply to query different sources to determine whether a variant is present in the dataset or not, without revealing everything about the data. If more or different data is required to complete the analysis, further requests are made to the data access committees. The work of the GA4GH will pioneer a way to enable more global sharing of publicly accessible genomic data. As part of the G4GH, the Genomic Data Working Group (http://genomicsandhealth.org/our-work/working-groups) focuses on data representation, storage, and analysis, including working with platform development partners and industry leaders to develop standards that will facilitate interoperability. It works closely with the G4GH Regulatory and Ethics Working Group to ensure that any data access or sharing complies with ethical consent associated with the datasets.

Second, the global effort to use genomic approaches to solve human health problems needs to be more equitable in terms of inclusiveness of ancestrally diverse study populations. Given what is known about the existence of population specific disease genetic variants and the value of trans-ethnic mapping for discovery, replication and fine-mapping, it is a scientific imperative that genomics becomes more globally diverse. Currently, over 90% of all GWAS focused on people of European ancestry, leaving the promise of genomic science unlikely to be realized in a timely and efficient manner for all human populations. With the funding of major international initiatives, such as the Human Heredity and Health in Africa (H3Africa)[1] and the establishment of the Mexico National Institute of Genomic Medicine (INMEGEN), the global community is beginning to address this problem. With over $70 M multiple year commitment from the NIH and the Wellcome Trust, H3Africa is creating and supporting a pan-continental network of labs, bioinformatics and

☆ The views in this article are those of the authors and do not represent those of their institutions. The authors declare no conflicts of interests.

* Corresponding authors.
E-mail addresses: rotimic@mail.nih.gov (C. Rotimi), Nicola.Mulder@uct.ac.za (N. Mulder).

[1] Rotimi C., et al. H3Africa Consortium. Research capacity. Enabling the genomic revolution in Africa. Science. 2014 Jun 20;344(6190):1346-8.

http://dx.doi.org/10.1016/j.atg.2014.09.002
2212-0661/

biorepository facilities that are applying leading-edge genomic and epidemiologic approaches to study the genetic and environmental bases of disease susceptibility and drug responses in Africans. H3Africa is already changing the landscape of genomic research in Africa.[1] INMEGEN was established by the Mexican congress in 2004 as a means of finding new strategies to tackle common diseases. This effort which includes the construction of a haplotype map of the Mexican populations and the development of large genetic epidemiology projects for the study of common diseases and biomarker discovery is beginning to provide novel insights into the history and health of the Mexican people. We strongly advocate for the expansion of these global efforts in low resourced nations with the goal of establishing state-of-the-art genomic infrastructure and training. We envision that this goal can be achieved by first understanding the political, economic and scientific expertise needed across national and international interest groups including local government agencies, private biomedical funding agencies, healthcare providers, biotechnology companies and advocacy groups among others. Second, trusted partnership with the strongest ethics and privacy requirements must be developed. Third, these partnerships must acknowledge and be willing to work to overcome the challenges of some its members such that these global arrangements are not perceived simply as a way to make valuable genomic and clinical data available to scientists in the industrialized nations. If concerns are addressed properly, it is indeed possible that these global collaborations will ensure the full participation of ancestrally diverse populations in genomic science as scientists and study participants with the ultimate goal of avoiding the exacerbation of current inequities in health and lack of diversity in biomedical research.

Lastly, we advocate for global data and knowledge sharing policies that recognize the strengths and challenges of scientists in different economic and cultural settings. The process of generating genetic epidemiology datasets usually requires costly investment of time and effort on the part of several scientists. In this regard, it is important to derive as much benefit as possible, in terms of publications and training of students in genomics and other aspects of biomedical research. However, this should not be at the expense of global data sharing. To ensure that the efforts of scientists generating the original data are adequately recognized, it is important that they have the necessary infrastructure and scientific capacity to take a timely advantage of the data that they have generated. Achieving this goal may be problematic for young or new investigators to the field of genomics and in general for scientists from developing countries who do not have access to extensive resources available to some researchers in the developed world. This highlights the need to accelerate the training of these investigators and to equip them with the infrastructure required through injection of funds such as those provided within the H3Africa and other similar initiatives.

Lack of access to skills and infrastructure would lengthen the time for researchers to achieve their research objectives, and if their data becomes publicly accessible too soon, they will lose the opportunity to be first to publish on the data. Processing of large volumes of genomic data, for example, will take much longer where computing resources are limited, and even transferring the data from the site of generation (e.g. sequencing centers) to the site of analysis is a major challenge in developing countries due to poor internet connectivity. This can be addressed by publication embargo policy that recognizes these limitations by providing relatively long time to publish their papers before making the datasets publicly available. A recent example is the data sharing policy of the H3Africa initiative available at www.h3africa.org. The H3Africa consortium data sharing and access policy provides for an extended time period for analysis prior to submission to public repositories. Concessions by funding agencies that take into account available resources of groups they fund will be important for protecting the interests of these groups, but ultimately, the data should be shared to enable others to benefit from the data and to use it for the advancement of science, economy and health of global populations. In all, data sharing across national and institutional boundaries is good for science and for the creation of “global good”. However, sharing has to be done ethically and fairly with a clear process of community engagement when appropriate and a robust process of obtaining individual informed consent for both primary and secondary uses of generated data.