Positioning and Power in Academic Publishing: Players, Agents and Agendas
F. Loizides and B. Schmidt (Eds.)

doi:10.3233/978-1-61499-649-1-48

Genome sharing projects around the world: how you find data for your research

Fiona NIELSEN[a,b,1] and Nadezda KOVALEVSKAYA[a,b]

[a] *Repositive Ltd, Future Business Centre, Kings Hedges Road, Cambridge CB4 2HY, United Kingdom*

[b] *DNAdigest, Future Business Centre, Kings Hedges Road, Cambridge CB4 2HY, United Kingdom*

Abstract. Access to raw experimental research data and data reuse is a common hurdle in scientific research. Despite the mounting requirements from funding agencies that the raw data is deposited as soon as (or even before) the paper is published, multiple factors often prevent data from being accessed and reused by other researchers. The situation with the human genomic data is even more dramatic, since on the one hand human genomic data is probably the most important data to share - it lies at the heart of efforts to combat major health issues such as cancer, genetic diseases, and genetic predispositions for complex diseases like heart disease and diabetes. On the other hand, since it is sensitive and personal information, it is often exempt from data sharing requirements. DNAdigest investigates the barriers for ethical and efficient genomic data sharing and engages with all stakeholder groups, including researchers, librarians, data managers, software developers, policy makers, and the general public interested in genomics. Repositive offers services and tools that reduce the barriers for data access and reuse for the research community in academia, industry, and clinics. To address the most pressing problem for public genomic data: that of data discoverability, Repositive has built an online platform (repositive.io) providing a single point of entry to find and access available genomic research data.

Keywords. Genomic data, data access, data sharing, genomic data repositories, tools

1. Introduction: data access and reuse is a common hurdle in scientific research

Research organisations, both public and private, are producing ever increasing volumes of data. Irrespective of whether the research is funded publicly or privately, there is increasing pressure to provide evidence that the maximum benefit is obtained from generated data. Recent years have seen a concerted effort by providers of public funds for research to require that the results of that research be publicly available (Collection of UK funders' policies 2015).

While the benefits of data sharing are becoming more widely accepted (Toronto International Data Release Workshop Authors 2009), human genomic data (i.e.,

[1] Corresponding Author: fiona@repositive.io

information about the composition of our DNA and RNA) is often exempt from data sharing requirements from major funders that all experimental data must be placed in publicly accessible repositories. This is because of concerns that making human genomic data public exposes potentially sensitive personal information to the world (Richards 2015).

2. The special case of human genomic data: it is there but it is mainly inaccessible

It is estimated that, in 2015, the world human genome sequencing capacity will exceed 80 petabyte of sequence a year. However, as of 2014, the largest public repository for human genomics data (the NIH database of genotypes and phenotypes dbGaP) holds only about 0.5 petabytes of clinical genomics data.

This gap between the availability of genomic information and the production of it can be at least partially attributed to the absence of tangible benefits for the individuals who make data available and, at the same time, to the existence of sanctions for improper handling of personal information. However, when data donors give consent for their data to be used for research, they set their expectation that the data will actually be used for this purpose. To not utilise their data in the best possible way within the consent given goes against the data donor's interests and expectations. Ironically, human genomic data is probably the most important data to share, since it lies at the heart of efforts to combat major health issues such as cancer, genetic diseases, and genetic predispositions for complex diseases like heart disease and diabetes. In particular, the promise of personalised medicine (where treatment is tailored to the individual) is unlikely to be realised without widespread access to large amounts of genomic data.

Existing data sharing initiatives generally take the form of some kind of repository for storing data or some kind of service to help find collaborators or data. Examples of repositories include publicly funded repositories (e.g. SRA, ENA, dbGaP, EGA, ArrayExpress etc), biobanks, and data repositories set up by individual institutions or projects (e.g. LOVD). Examples of services to help find data include, for example, GenomeConnect, PhenomeCentral and the Beacon project.

Public data repositories do an excellent job of storing data, a crucial task to enable data availability. The mentioned services do great jobs at servicing specific needs, e.g. connecting clinicians who have found similar phenotypes for their patients with genetic diseases. Currently no public initiatives are have successfully addressed the problem of discovering the existence of datasets (data discoverability) for specific diseases and specific data types across locations and repositories. In addition, all of the mentioned initiatives face challenges of funding and sustainability of their initiatives, due to their reliance on research grants.

3. A solution to increase access to human genomic data: the community platform

To address the most pressing problem for public genomic data: that of data discoverability (van Schaik 2014), Repositive has built an online platform

(repositive.io) providing a single point entry to search and access public genomic data sources. The Repositive platform enables users to search through all its indexed data sources in a single click via an easy-to-use interface free of charge. To address the problem of varying quality and type of metadata associated with data across data sources and public repositories, the Repositive platform allows users to comment on the content and quality of datasets and add descriptions to the listed metadata. If a research scientist has data that he/she would like to share but cannot for any reason, he/she can announce the existence of the data on the Repositive platform. In this case, other scientists that have similar or complementary data can contact the author to start a collaboration or to discuss for instance the conditions under which they can exchange their data. Similarly, a user can post a request for data and another user, who has the data stored but not used, can respond and find an application for their otherwise unused data.

By listening to our users and concentrating on a specific use case for the genetics researcher – the problem of finding and accessing human genomic research data - and supporting best practices for data annotation, accessibility and reuse, we offer the Repositive platform and services as a contribution to ease the workflow for research in human genomics for health and disease.

The Repositive business model is built around the Repositive freemium features of the online platform for data discovery. The online platform is open for all to sign up and search for free (see above), but at the same time Repositive offers premium products and services to both data providers and data consumer organisations. Our premium services include: customised data scouting; data access applications; automating data access workflows; setting up and maintaining public data catalogues; setting up data collaborations between different organisations, e.g. across industry and academia, etc. With this business model, Repositive can deliver a free service on the online platform which supports researchers across academia and industry, while our revenue comes from related professional products and custom services.

References

Collection of UK funders' policies. In: Research Data Management Blog. (2015) Available at: http://www.data.cam.ac.uk/funders. (Accessed: 15 February 2016).

Richards, M., Anderson, R., Hinde, S., Kaye, J., Lucassen, A. et al. (2015) 'The collection, linking and use of data in biomedical research and health care: ethical issues'. Nuffield Council on Bioethics. Report. Available at: http://nuffieldbioethics.org/wp- content/uploads/Biological_and_health_data_web.pdf. (Accessed: 15 February 2016).

van Schaik, T.A., Kovalevskaya, N.V., Protopapas, E., Wahid, H. and Nielsen, FGG (2014) 'The need to redefine genomic data sharing: A focus on data accessibility'. Applied&Translational Genomics, 3(4), pp.100-104 (doi:10.1016/j.atg.2014.09.013)

Toronto International Data Release Workshop Authors (2009) 'Prepublication data sharing'. Nature, 461, pp. 168-170. (doi:10.1038/461168a)