Data Governance in the Genomics Era

Emma Lawrence, Department of Plant Science, University of Cambridge

In recent years, the volume of data generated from all aspects of our lives has been increasing, in parallel with the sophistication of analytical techniques used to process this data. This shift toward a 'data-driven' society has the potential to yield insights that can benefit many sectors of public life, but it has also prompted concerns related to privacy. A recent report by the Royal Society on data management and use [1] is a recognition that the fast pace of all areas of data growth requires careful consideration.

In the field of healthcare research, an area generating large amounts of highly unique data about individuals is that of genome sequencing and genomics. Sharing of genome sequence data has the potential to improve our understanding of diseases, which can, in turn, improve diagnostics, treatment and integration of personalised medicine into standard healthcare practices. However, the difficulties associated with maintaining privacy of this data are significant. These challenges demand a need for policies that will encourage innovation and scientific progress for the collective benefit of all whilst minimizing the level of risk to the individual.

This short article will explore the potential advantages and risks of using genomic data in medical research, and it will suggest policy approaches to address these challenges.

What is genomics and how can it be used for healthcare?

The DNA of all organisms is composed of a long sequence of DNA nucleotides – A, C, T and G – that together form a unique code. Through genome sequencing, scientists can determine the order of these letters in an individual organism. All humans have the same nucleotide letter for most positions in the genome, but they differ at a few positions which are termed 'variants'. While most variants in the genome do not impact our physiology, some can cause disease. Knowledge of these variants can be useful for informing treatment, as well as for providing timely diagnoses. Many of these disease-causing variants are rare, meaning that they are not observed at high rates in the general population. As such, genomic analysis requires large datasets comprised of many—typically thousands—of genome sequences, so researchers have enough statistical power to detect such variants. Luckily, the cost of sequencing a genome has plummeted in recent years, and therefore many individuals can be sequenced synchronously for minimal costs. Nonetheless, data sharing, which is simply the combining of different smaller datasets generated in different research centres, can help produce the large datasets required. It can also increase efficient interpretation of the same variants across different research centres, reduce the risk of misdiagnosis, and improve the reliability of diagnoses [2]. Taken together, data sharing can be of a direct benefit to patients living

with rare diseases, and the UK has adopted several policies to encourage further data pooling [3]. Genomics England is leading a movement to adopt genomic testing as an integrated part of routine clinical care in the NHS, and the ongoing 100,000 Genomes Project aims to set up a genomic medicine service for the NHS in the coming decades [4].

Concerns and Risks

We have seen that the collection and sharing of genomic data has the potential to bring advances in scientific understanding and healthcare. However, there are some concerns associated with this.

First, guaranteeing the privacy of individual-level genomic data can be challenging. Data shared between research groups is typically 'de-identified', meaning that any personally identifiable information (PII) must be removed from the dataset before genomic data can be shared with other research groups. While PII most obviously includes information like name, date of birth and home address, other information, such as a post code, county or even ethnicity, could be combined with other PII to identify an individual, particularly those with rare diseases. In the case of patients with these diseases, there is a concern that a breach of confidentiality of this information could place them at risk of being subject to discrimination and/or stigmatisation. However, the de-identification of data could limit the ability of researchers to contact an individual in the future, for example if they are thought to have increased risk of a disease [2]. This can be circumvented by using 'coded data,' so individuals can still be linked to their genomic data and identification can occur if required, but the code is kept in a secure environment. However, it has been suggested that DNA can never be completely anonymised due to the inherent uniqueness of the genetic identity [5]. Current legislation does protect and regulate the sharing of personally identifiable data, but there is a lack of consensus over the appropriate level of safeguarding for genomic data to minimize privacy risks.

A concern for the collection of genome data is how to obtain consent for its usage. An individual may consent for their own personal genome being sequenced and the data released, but this can also give indirect information about family members, and to a lesser extent, members of the same ethnic group and population [1, 6]. Therefore, some question whether a genome sequence can be 'owned' and consequently whether one individual can consent to its use. It is also difficult to consent to all the possible future uses of the data. Both data analysis and genomics are rapidly advancing fields, and it may not be possible to foresee all future possibilities. A 'broad consent' model permits use of the data for an unspecified range of future research in recognition of these difficulties, but it is important that individuals understand what this consent means in practice.

Suggested policy approaches

Several different sources have argued for new regulatory bodies to address the challenges of a changing genomic medicine landscape. The Science and Technology Committee recently launched an inquiry into genomics and genome editing, where suggestions were made that a new body, similar to the Human Genetics Commission which existed up to 2012, should be formed [7]. In her 2016 annual report 'Generation Genome', Dame Sally Davies recommends that government public engagement with genomics should be increased with the creation of a new National Genomics Board [5]. This approach will help to ensure that progression will be monitored and investigation into any potential harm is carried out.

A consensus for how genomic data will be

confidentially treated should be reached. If successful, lessons can be taken from the 100,000 Genomes Project and applied to other projects. They have created a secure data governance system for storage and access of sensitive patient data, where de-identified data is analysed in a monitored environment. Researchers need to apply to access the de-identified data which can only be approved if the purpose is deemed reasonable. In addition, the database of Genotypes and Phenotypes, which is a National Institutes sponsored repository of large-scale genetic and clinical datasets, has a rigorous application process for annonymised data and requires research institutes to provide secure data storage that aligns with their guidelines [8]. In agreement with this, a report by the Nuffield Council on Bioethics also makes the following recommendations; that privacy breaches must be reported to affected individuals, that criminal penalties should apply for misuse of data, and that access to data is restricted to researchers that are subject to institutional oversight [3].

Another consideration is the importance of cultivating public trust in any genome sequencing project. As in any area of human subjects research, the security of data storage must be made fully transparent to those involved in a study, and researchers should acknowledge that privacy cannot be completely guaranteed. An example of a healthcare data project that failed because it did not cultivate public trust was the NHS's care.data program. The purpose was to extract data from GP practices and link it with that from hospitals, to improve treatments and patient care. However, it was stopped in 2016 after concerns over data privacy weren't fully addressed or communicated to patients [9,10]. Despite extensive patient communication and public dialogue, there remains confusion over the concept of anonymised and pseudo-anonymised data in the 100,000 genomes project [11]. This highlights the importance of maintaining a clear dialogue with the public. Finally, new uses for genomic data emerge every year, and policymakers should consider how obtaining informed consent at each stage of these new developments could increase an individual's knowledge and ownership over the use of their data.

Conclusions

It is expected that as genome sequencing and genomic testing becomes more commonplace in research and healthcare, a shift in the policy landscape will be required to manage the associated risks. It is important that scientific progress in this area can continue, but in a secure environment that people trust. Public participation is vital for the success of future genomic research projects, and their promise to deliver transformative genomic medicine.

Acknowledgements

First editor: Maggie Westwater
Second editor: Erin Cullen
Image credit: Shaury Nash, CC BY-SA 2.0

References

[1] The Royal Society (2017) *Data management and use: Governance in the 21st century*. [Online]. Available: http://bit.ly/2C9TzCT

[2] Raza, S., and Hall, A. (2017). Genomic medicine and data sharing. *Br. Med. Bull. 123*, 35–45.

[3] Nuffield Council on Bioethics. *The collection, linking and use of data in biomedical research and health care: ethical issues.* [Online]. Available: http://bit.ly/16qU2A4

[4] Genomics England: The 100,000 Genomes Project [Online]. Available: http://bit.ly/2bbfJuG

[5] Chief Medical Officer Annual Report 2016: Generation Genome. [Online]. Available: http://bit.ly/2uhbNQW

[6] Royal Society Talk by Chloe-Agathe Azencott (2017). *Machine learning and Genomics: precision medicine vs patient privacy.* [Online]. Available: http://bit.ly/2pdNgJK

[7] House of Commons Science and Technology Committee (2017) *Genomics and genome-editing: future lines of enquiry.* [Online]. Available: http://bit.ly/2IrkywO

[8] Mailman, M.D. et al., (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet. 39, 1181-1186.*

[9] Limb, M. (2016). Controversial database of medical records is scrapped over security concerns. *BMJ 354*, i3804.

[10] Godlee, F. (2016). What can we salvage from care.data? *BMJ 354*, i3907.

[11] The Guardian (2015) *Privacy and the 100,000 Genome Project.* [Online]. Available: http://bit.ly/2HEzZAq

About the Author

Emma is a PhD Student in the Department of Plant Sciences, working on meiotic recombination in the flowering plant *Arabidopsis thaliana.* Emma completed her undergraduate degree in Natural Sciences at the University of Cambridge, where she developed a passion for genetics and the complexity of genomes. She is interested in the role of policy in promoting genomic medicine, food security, and the use of evidence-based policy in government. During her PhD, she completed an internship with Sense about Science in London.