COMPUTER LAW & SECURITY REVIEW 000 (2018) 1–19

Available online at www.sciencedirect.com

ScienceDirect

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

Computer Law & Security Review

# Big genetic data and its big data protection challenges

*Paul Quinn*[a,*], *Liam Quinn*[b]

[a] *Vrije Universiteit Brussel, Brussels, Belgium*
[b] *University College London, United Kingdom*

ARTICLE INFO

*Article history:*
Available online xxx

*Keywords:*
Genetic research
Big data
Data protection
GDPR

ABSTRACT

The use of various forms of big data have revolutionised scientific research. This includes research in the field of genetics in areas ranging from medical research to anthropology. Developments in this area have inter alia been characterised by the ability to sequence genome wide sequences (GWS) cheaply, the ability to share and combine with other forms of complimentary data and ever more powerful processing techniques that have become possible given tremendous increases in computing power. Given that many if not most of these techniques will make use of personal data it is necessary to take into account data protection law. This article looks at challenges for researchers that will be presented by the EU's General Data Protection Regulation, which will be in effect from May 2018. The very nature of research with big data in general and genetic data in particular means that in many instances compliance will be onerous, whilst in others it may even be difficult to envisage how compliance may be possible. Compliance concerns include issues relating to 'purpose limitation', 'data minimisation' and 'storage limitation'. Other requirements, including the need to facilitate data subject rights and potentially conduct a Data Protection Impact Assessment (DPIA) may provide further complications for researchers. Further critical issues to consider include the choice of legal base: whether to opt for what is often seen as the 'default option' (i.e. consent) or to process under the so called 'scientific research exception'. Each presents its own challenges (including the likely need to gain ethical approval) and opportunities that will have to be considered according to the particular context in question.



## 1. Introduction

The use of genetic data in research has been undergoing a fundamental shift. Researchers are no longer restricted to working with relatively small samples of individual genomes (for example DNA relating to a gene known to effect disease aetiology) but now work with various markers scattered across the entire genome. This type of data is used in various areas of research including efforts to discover new disease variants or to increase understanding of evolutionary processes. The field of bioinformatics and computational genetics has evolved inter alia to allow researchers to focus on detailed 'high-depth' sequencing of the entire genome of individuals allowed by advances in genome sequencing technology and computing power. These advances mean that an individual's genome can be sequenced relatively quickly and cheaply (costing less than a MRI scan in a local hospital). Powerful software has

* Corresponding author: Law Science Technology & Society (LSTS), Vrije Universiteit Brussel, Building B, room 4B31, Pleinlaan 2, B-1050 Brussels, Belgium.
*E-mail address:* paul.quinn@vub.be (P. Quinn).

https://doi.org/10.1016/j.clsr.2018.05.028
0267-3649/

Please cite this article as: P. Quinn, L. Quinn, Big genetic data and its big data protection challenges, Computer Law & Security Review: The International Journal of Technology Law and Practice (2018), https://doi.org/10.1016/j.clsr.2018.05.028

furthermore been developed to analyse such genome wide sequences (GWSs). The research potential of such techniques has been complimented by the ability to share and combine GWS data with a range of potential complimentary data sets (e.g. electronic health records). These developments have ushered in a world of 'big data genomics' where researchers carry out complex data mining operations on the entire genomes of individuals and groups of individuals.

Whilst these developments promise to permit great leaps forward in our understanding of the human genome and its relationship to various important issues (not least to human disease), they also pose new risks in terms of privacy related harms. These include harms not only to the individuals providing the genetic samples in question but even to those who may be related to them.[1] Complying with laws relating to privacy, and in particular to data protection will therefore be a serious issue for researchers conducting research on large samples of genetic data. This article aims to illustrate a number of these issues, highlighting some of the major challenges that the data protection framework poses for researchers active in the use of big genetic data.[2] It will focus on compliance with the EU's new General Data Protection Regulation (GDPR), which comes into effect across the EU from May 2018. In doing so this paper will use several prominent examples from documented research practice in the area of computational genetics. The authors will illustrate how common practices in this area may be difficult to reconcile with the key pillars of data protection, including the need to have a valid legal ground for processing personal data, the need to respect data processing principles and the need to facilitate data protection rights. As this paper suggests, such burdens may mean that compliance with the EU's data protection regime (including under the new General Data Protection Regulation) may not only be cumbersome but may, in many cases, be difficult even to envisage given the aims of big genetic data processing for research.

Section 2 of this paper will briefly introduce the concept of 'big genetic data' and discuss how researchers can use it. Sections 3 and 4 will look at how, given the nature of modern computational genetics', genetic data used in research is likely not only be to be of a personal nature, (i.e. rarely anonymous in nature) but also categorised as 'sensitive' or 'special' data also. Section 5 will look at how the need to respect data processing principles will present difficulties for researchers involved in computational genetics. Section 6 will look at the issue of data protection impact assessments, something that will be obligatory (and potentially onerous) for many forms of research given the sensitive (or special) nature of genetic data. Section 7 will analyse how the need to facilitate data subject rights may create major obstacles for researchers involved in the use of big genetic data. The issues surrounding the use of both consent and the scientific research exception as a legal base for processing will be discussed in Sections 8 and 9 respectfully. The requirements of each may mean that on many occasions the latter is more suitable, though as Section 9 discusses this may be something researchers (including in areas of computational genetics) have difficulty in convincing ethics committees of, presenting further problems for research in this area.

## 2. Big genetic data and its use in research

Genetic data originates from human tissue or other biological samples. These range from blood, saliva and urine samples taken from individuals to tissues taken from cadavers in ancient DNA studies to soil, water and rock samples in environmental DNA studies.[3] DNA is a double stranded nucleic acid molecule found in the nucleus of nearly all cells in the human body. It is a ladder shaped molecule composed of two sugar-phosphate backbones linked by nitrogenous bases of which there are four types. It is the order or sequence of these bases that gives rise to genetic code. In order to sequence DNA it has to be separated from the surrounding medium it is contained in and then purified from other cellular components using various laboratory techniques.[4] In cases where a minuscule amount of DNA is obtained (often the case in forensic science), the DNA is then amplified using various biochemical techniques to produce a sufficient amount for sequencing purposes.

Different genetic projects vary greatly in the number and type of genetic data that is collected, processed, stored and disseminated to other researchers and research groups. One general trend has been that the sample size (number of participants) and the amount of genetic data that researchers work with has increased enormously in the recent years. Indeed, the use of Genome Wide Samples (GWSs) is becoming increasingly common. This is unlike earlier research that may have involved a limited portion of the genome. In addition, researchers may seek to combine GWSs with various forms of complimentary data that aid analysis.[5] This may include disease status, age, geographical origin, and various other measures. Such measures allow researchers to track relationships and patterns between certain variables and DNA sequences.

[1] See for example: Nuffield Council on Bioethics. The Linking and Use of Biological and Health Data; 2013. Available online at: http://nuffieldbioethics.org/wp-content/uploads/Biological_and_health_data_web.pdf.

[2] In doing so it draws on the expertise of the authors, one of whom is a specialist in data protection law and health data in particular, the other is a specialist in computational genetics.

[3] Lugg, W., Griffiths, J., Van Rooyen, A., Weeks, A. & Tinglet, R. 2017. Optimal survey designs for environmental DNA sampling. *Methods in Ecology and Evolution,* DOI: 10.1111/2041-210X.12951. Livy, A., Sayhean, L., Jagdish, C., Hanis, N., Sharmila, V. & Wee Ler, L. P., B 2012. Evaluation of Quality of DNA Extracted from Buccal Swabs for Microarray Based Genotyping. *Indian Journal of Clinical Biochemistry,* 27, 28-33. Deribe, K., Beng, A., Cano, J., Njouendo, A., Fru-Cho, J., Awah, A., Eyong, M., Chounna Ndongmo, P., Giorgi, E., Pigott, D., Golding, N., Pullan, R., Noor, A., Enquselassie, F., Murray, C., Brooker, S., Hay, S., Enyong, P., Newport, M., Wanji, S. & Davey, G. 2018. Mapping the geographical distribution of podoconiosis in Cameroon using parasitological, serological, and clinical evidence to exclude other causes of lymphedema. *PLOS Neglected Tropical Diseases,* https://doi.org/10.1371/journal.pntd.0006126.

[4] Butler, J. 2015. The future of forensic DNA analysis. *Philosophical Transactions of the Royal Society,* DOI: 10.1098/rstb.2014.0252.

[5] Nuzzo, A., Riva, A. & Bellazi, R. 2009. Phenotypic and genotypic data integration and exploration through a web-service architecture. *BMC Bioinformatics,* https://doi.org/10.1186/1471-2105-10-S12-S5.

Please cite this article as: P. Quinn, L. Quinn, Big genetic data and its big data protection challenges, Computer Law & Security Review: The International Journal of Technology Law and Practice (2018), https://doi.org/10.1016/j.clsr.2018.05.028