ARTICLE IN PRESS
JID: CLSR [m7;July 13, 2018;11:11]
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/

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

As an example many common genetic association studies attempting to link certain DNA polymorphisms (DNA bases that vary in a population) with traits (e.g. disease and non disease) will utilise huge genetic databases collected for widespread use and with accompanying phenotypic data including height, BMI, smoking status, alcohol consumption, medical history and even ancestral genealogical history.[6] The ability to generate a link between a genetic polymorphism and a trait in these types of study will depend on having access to accurate genetic and phenotypic data, the heritability of the trait (genetic component in explaining trait variation) and the statistical power in the analysis (influenced heavily by the number of samples available to the researcher). Examples of prominent genetic databases created for research purposes include the UK10K project and the 100,000 genomes project.[7] These are both large consortium based research projects with the intention of providing huge datasets to different research groups. The UK10K project is an attempt to better understand causes and discover variants in relation to rare genetic diseases. The 100,000 genomes project focuses on cancer and rare diseases. Other uses of such data may occur in projects concerned with developing a better understanding of human evolution and migration. Given not only the use of GWS, which by themselves are heterogeneous data sets of enormous size, but also the combination with other potential data sources researchers are effectively making use of what can be considered 'big genetic data'.[8]

3. Is big genetic data always personal data?

3.1. *It is becoming easier to link genetic data to specific individuals*

Personal data is data that can likely be linked to an identifiable individual. Data that cannot be linked to an individual is not personal data and is not governed by the EU data protection framework.[9] Consequently, those involved in processing such data will not have to comply with its requirements. Where possible, researchers have in the past tended to claim that genetic data was not personal data in order to avoid the need for compliance with data protection regulations. This was particularly the case (arguably justifiably) where it had often been assumed that data related to certain limited sequences of the genome could be considered as being anonymous in nature if it was not stored with any information that would link it directly to an identifiable individual (e.g. name, date of birth, unique identifying numbers etc.).[10]

Developments in computational genetics, in particular the use of big genetic data make such assumptions obsolete in many cases. These developments can be broadly categorised in three ways. *First*, the ability of researchers, inter alia through the World Wide Web and associated forms of connectivity, to share and access data around the world has increased enormously. This includes not only the ability to access and share genetic sequences but also complimentary forms of data. *Second*, computing power has continued to grow at a pace envisaged by Moore's law. This has inter alia permitted the acceleration of the *third* important factor, i.e. the development of ever more powerful algorithms that are capable of more thorough data analysis and, in a number or contexts, allowing identification of individuals where it was previously not thought possible. These three developments together (as the practical examples below demonstrate) mean that it is increasingly possible to identify individuals from what might seem like an apparently anonymous sample of genetic code through the use of powerful statistical techniques and other forms of publically available data (including data from previous research projects). The potential likelihood of re-identifying individuals from apparently anonymous genetic data has increased given the common use of GWSs. The size of such samples means both that there is more material for processing algorithms to work with (thus increasing their power) and a greater likelihood that samples can be matched with externally available public data.[11]

3.2. *A high legal bar has been created for achieving anonymisation of data*

In addition to the technical factors described above, the threshold of what actually constitutes anonymous data has been set extremely high in legal terms. The article 29 working party has confirmed this in its analysis of what the term "reasonably likely" means in Directive 95/46/EC (this term is used to discern whether data is anonymous or not and refers to the possibility to de-anonymise the data in question).[12] In particular, it has outlined four factors that must be taken into consideration.[13]

First, it requires data controllers to focus on the means that would be necessary to bring about deanonymisation.[14] This requires a consideration of the ever-evolving technical possibilities in terms of computing power and the availability of

[6] Dubois, L., Kyvik, K., Girard, M., Tatone-Tokuda, F., Pérusse, D., Hjelmborg, J., Skytthe, A., Rasmussen, F., Wright, M., Lichtenstein, P. & Martin, N. 2012. Genetic and Environmental Contributions to Weight, Height, and BMI from Birth to 19 Years of Age: An International Study of Over 12,000 Twin Pairs. *PLoS ONE*, https://doi.org/10.1371/journal.pone.0030153.

[7] Consortium, T. U. K. 2015. The UK10K project identifies rare variants in health and disease. *Nature*, 526, 82-90. See also The 100,000 Genomes Project Protocol (2017) v3, Genomics England . doi: 10.6084/m9.figshare.4530893.v2.

[8] He, K. Y., Ge, D. & He, M. M. 2017. Big Data Analytics for Genomic Medicine. *International Journal of Molecular Sciences*, 18, 412.

[9] GDPR recital 26.

[10] Roewer, L. 2013. DNA fingerprinting in forensics: past, present, future. *Investigative Gentics*, https://doi.org/10.1186/2041-2223-4-22.

[11] Niemiec, E. & Howard, H. 2016. Ethical issues in consumer genome sequencing: Use of consumers' samples and data. *Applied Translational Genetics*, 8, 23-30.

[12] Article 4 of the GDPR uses a similar description.

[13] Quinn, P. 2017. The Anonymisation of Research Data — A Pyric Victory for Privacy that Should Not Be Pushed Too Hard by the EU Data Protection Framework? *European Journal of Health Law*, 24, doi 10.1163/15718093-12341416.

[14] Khaled, E. & Alvarez, C. 2015. A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques. *International Data Privacy Law*, 5, 73-87.p75.

algorithms that are able to deanonymise data that was thought to be anonymous.[15] *Second*, it is necessary to take into account the fact that many types of publicly available datasets that are claimed to be anonymous may not meet the requisite standards of anonymisation.[16] Such a standard requires that the party creating the anonymising data, not only consider their own ability to deanonymise the data in question but also the ability of other known and unknown parties given the state of technological development and other potential sources of data that may be publicly available. Imagine for instance the use of genetic data or its publication online following research (something, which as Section 3.3 describes, is usually seen as good research practice).[17] Given the nature of the data involved and the potential for related information concerning the individual or a family member to exist in an accessible version elsewhere, it may be difficult to speak of genetic data as ever being truly anonymous.[18] *Third*, in assessing whether a state of actual anonymisation exists, one cannot depend upon the "good motives of the data controller".[19] This means that the data controller, when assessing whether a dataset he or she possesses is truly anonymous, must take into account what other data they have access to. If the controller of the supposedly anonymised dataset has access to other data that will allow the identity of individuals to be discerned through cross-referencing the two, then it is not correct to speak of anonymised data. *Fourth*, the working party confirmed that in its opinion the act of anonymisation itself constitutes an act of processing of personal data.[20] This is logical, as in order to anonymise data the data controller must have been in possession of data that was not anonymised i.e. personal data. As *Kaheld* states, "In order to anonymise data, it is necessary for an anonymisation engine to ingest personal data, apply anonymisation techniques to it, and then output anonymised data. The input is personal data." [21] Given this, it is also logical to expect that the dataset in question be obtained in accordance with one of the legal bases described in Sections 8 and 9. This may create a "catch 22" because it means that in order to collect personal data, even if the intention was to immediately anonymise it, it would be necessary to have the consent of the data subjects involved or possesses some other valid legal base for such processing (see Sections 8 & 9). Where the purpose of anonymisation is to avoid the need to obtain consent, this will present immediate problems because, where such consent has not been obtained, it may not be possible to gather the data in the first place.

In addition to the suggestions of the Article 29 Working Party, one should also take note of the guidance given by the European Court of Justice (CJEU). This particularly includes the judgement *Breyer*, (Case C-582/14: *Patrick Breyer v Bundesrepublik Deutschland*). In responding to a referral from the German Federal court, the CJEU clarified how the term 'reasonably likely' should be understood. In doing so, the court outlined an expansive notion that arguably reduces the scope for claiming that genetic big data is not personal data. It ruled that data could not be thought of as anonymous if a potential controller could, through efforts that were not disproportionately difficult or illegal, obtain further data which, by combining with the data available, would allow identification of the data subject(s) in question. Such factors are important in the context of big genetic data research projects given that there may be a wealth of information that is publicly available, legal to access and which would not require an unreasonably disproportionate effort to obtain. As Section 3.3 discusses, this includes research data made public by other research projects and data to which researchers may have access to in order to conduct scientific research such as EHRs. There may furthermore be a wealth of complimentary data freely available online. Given that researchers may be able to access such data without disproportionate effort and through legal means (accessing and processing data that has been manifestly made public is for instance permitted under GDPR Article 9(2)(e)) the likelihood that big genetic data alone can be thought of as non personal nature is arguably greatly reduced.

Taking into account the cumulative effect of such requirements and the nature of GWS analysis, the authors of this paper would argue that anonymisation of big genetic research data may no longer be considered realistic.[22]

3.3. *The potential threat of deanonymisation*

Most research projects do not process genetic data in isolation but in combination with other data relating to non-genetic factors (e.g. lifestyle variables or socioeconomic data).[23] Whilst this increases the power of statistical analysis for researchers, it also has a potentially problematic consequence: the more of this associated data that is combined with the genetic data the easier it is to identify an individual inter alia by linking it with some other available dataset. An individual's genome is fixed for life and cannot be altered. This means that genetic data will retain a permanent link with specific

[15] See Article 29 Working Party Opinion on Anonymisation p9.

[16] Many such datasets may more realistically be described as 'psedonymised'. For more see: Aldhouse, F. 2014. Anonymisation of personal data - A missed opportunity for the European Commission. *Computer Law and Security Review,* 30, 403–418.

[17] Mcguire, A. L., Hamilton, J. A., Lunstroth, R., Mccullough, L. B. & Goldman, A. 2008. DNA data sharing: research participants' perspectives. *Genet Med,* 10, 46-53.

[18] Schmidt, H. & Callier, S. 2012. How anonymous is 'anonymous'? Some suggestions towards a coherent universal coding system for genetic samples. *Journal of Medical Ethics,* 38, doi:10.1136/medethics-2011-100181. Bohannon, J. 2013. Genealogy Databases Enable Naming of Anonymous DNA Donors. *Science,* 339, doi: 10.1126/science.339.6117.262.

[19] See Article 29 Working Party Opinion on Anonymisation p10.

[20] Working Party Opinion on Anonymisaiton, p2: The working party states "Anonymisation constitutes a further processing of personal data; as such, it must satisfy the requirement of compatibility by having regard to the legal grounds and circumstances of the further processing."

[21] Khaled, E. & Alvarez, C. 2015. A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques. *International Data Privacy Law,* 5, 73-87. p79.

[22] Gymrek, M., Mcguire, A., Golan, D., Halperin, E. & Erlich, Y. 2013. Identifying personal genomes by surname inference. *Science,* 339, 321-324.

[23] Shoenbill, K., Fost, N., Tachinardi, U. & Mendonca, E. 2014. Genetic data and electronic health records: a discussion of ethical, logistical and technological considerations. *Journal of American Medical Informatics Association,* Doi: 10.1136/amiajnl-2013-001694.

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

individuals. This presents problems in terms of the need to make available research data for other researchers. Such requirements exist because a central pillar of science in general and in the life sciences in particular is replication of research findings.[24] In computational genetics, this is best facilitated by publication of genetic data online and allowing researchers at other institutions to download the data for their own analysis.[25] This may feasibly mean making such data available long after primary research has finished.[26] This permits inter alia a whole sector of 'secondary data researchers' to use pre-existing data to answer their own research questions as well. This allows these groups to avoid the cost and expense of gathering primary genetic data, safeguarding its storage and obtaining consent from participants.[27]

Historically, the usual approach for 'anonymising' genetic data that was made publically available was to remove any identifying information directly linked to the sample. This has included elements such as name, address or date of birth. Efforts may also have been made to remove any accompanying data that could have led to the sample being identifiable including health records, geographic data (including town, city and village of residence), and other phenotypic data that may lead to identification of an individual. Other administrative practices employed to make identification of a participant more difficult may have included the restriction of access to genetic data to only research groups that meet certain criteria and the creation of access agreements with the primary institution that gathered the data. Various other administrative hurdles relevant to the particular context in question may also have been employed to restrict access to data (and thus in theory reduce the chances of unauthorised identification of individual data subjects).[28]

Even in the context of the increasing use of such measures to anonymise data (all of which make research more difficult), there are concerns that anonymisation of genetic data is increasingly becoming an impossible task.[29] Such concerns become ever more important given the factors identified by the Article 29 working party as being relevant to the question of anonymisation (discussed above in (b)). Although it is clear that such measures will make the task of identifying an individual more difficult, concerns have mounted that the relentless increase of both computer power and algorithmic sophistication combined with the exponentially growing availability of potentially complimentary data may render such efforts largely futile. There have been a series of publications that have highlighted the possibility of identifying individuals from genetic data combined with other publicly available data in ways that may have been thought intuitively impossible. In one study, researchers were able, by comparing genetic data published by a research project to case control statistics in a published paper, to deduce whether the sample was from a person with the disease trait or a control sample.[30] In another study, the authors were able to combine the genetic data from one research project with publically available databases of surnames. The was possible given that Y chromosome data is known to correlate with surnames as it is inherited between generations on the paternal line only. After comparing the data in the publication to various genealogical databases online and various other online datasets, the authors were able specifically to name participants in the study.[31] Such developments highlight the fact that the possibility of identification is not only affected by what data and materials are made public by primary research groups but also by what is available online from a variety of other sources (e.g. ancestral websites, career networking sites, social societies) . Given that both computer/algorithm power and the availability of data will continue to grow in the future, the technical feasibility of considering large amounts of genetic data (and especially GWSs) to be anonymous will only become more dubious.

4. Personal genetic data is always sensitive data

Personal data that is sensitive in nature attracts a higher regulatory burden than non-sensitive data. The legal situation concerning genetic data is in a situation of flux. This is because the GDPR explicitly describes genetic data as 'special' (i.e. sensitive) data.[32] This was not the case with Directive

[24] There has been much discussion in the field of biology of a lack of replicability/reproducibility of findings that initially garnered much publicity and presented potential breakthroughs in understanding. Surveys have been conducted that highlight this concern within the field, in one instance 52% of respondents stated that there was indeed a 'crisis of reproducibility'. https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970.

[25] The 1000 Genomes Project is an example of a research project which has made its genetic data freely available to download for other research groups from an open access web portal. This data has been used in over a thousand different publications subsequently.

[26] Hudjashov, G., Karafet , T., Lawson, D., Downey, S., Savina, O., Sudoyo, H., Lansing, S., Hammer, M. & Cox, M. 2017. Complex Patterns of Admixture across the Indonesian Archipelago *Molecular Biology and Evolution,* 34**,** 2439-2452.

[27] Mcguire, A. L., Hamilton, J. A., Lunstroth, R., Mccullough, L. B. & Goldman, A. 2008. DNA data sharing: research participants' perspectives. *Genet Med,* 10**,** 46-53. See: Zheng-Bradley, X. & Flicek, P. 2017. Applications of the 1000 Genomes Project resources. *Briefings in Functional Genomics,* 16**,** 163-170.

[28] In the case of the 100kGP researchers have to attend the 100kGP information centre in Hinxton in order to access the dataset, administrators hope that this ensures that identifiable information remains physically contained in the primary institute and concerns about identification can be reduced.

[29] Hayden E, The Genome Hacker, Nature News (2013) Available at: https://www.nature.com/news/privacy-protections-the-genome-hacker-1.12940 In this article many scenarios are discussed whereby determined computer scientists are able to glean personal data from data systems and published research findings that were proclaimed to be protected or anonymised. The authors point to the growing difficulty and infeasibility of making such proclamations going forward given the growing proliferation of personal data available.

[30] Cai, R., Hao, Z., Winslett, M., Xiao, X., Yang, Y., Zhang, Z. & Zhou, S. 2015. Deterministic identification of specific individuals from GWAS results. *Bioinformatics,* 31**,** 1701-7.

[31] Gymrek, M., Mcguire, A., Golan, D., Halperin, E. & Erlich, Y. 2013. Identifying personal genomes by surname inference. *Science,* 339**,** 321-324.

[32] Article 9 GDPR.

95/46/EC. It did not define what genetic data was or what legal value it had. The Article 29 Working Party opinion on genetic data[33] described how Directive 95/46/EC is understood as implicitly allowing for genetic data to be recognised as a form of 'health data' which represents sensitive data and therefore subject to the application of stricter requirements (including, but not limited to the stricter requirements of explicit consent, where applicable). According to the working party, in order to demonstrate that genetic data was sensitive data it was necessary to show in a particular instance that such data was not only personal but also health data.[34] To do so it was necessary to show that the genetic data in question could provide an 'indication' as to the 'health status' of an identifiable individual.[35] This status represented an extremely wide potential reading of what 'health data' could include. This is because such a concept went beyond data that indicated the presence of disease by potentially including also data that could be used to infer the possibility (even if low) of disease developing but also even data that could simply confirm that an individual was 'healthy'.[36] Despite the evident breadth of the concept of health data, it is by no means a catch-all concept.

Although many forms of genetic data may be able to give an indication as to the health status of an individual not all such data will be able to. Historically this was arguably the case with many instances of genetic data. Even now, understanding of the functional role of the majority of the genome is still in its infancy. At the turn of the century, over 90% of the genome was frequently classified as 'junk DNA' (i.e. the relics of viral invasions and integrations into the human genome sequence and sequences that were previously functional but now redundant), something we know now is not the case.[37] The likely lack of any knowledge of any possible link between a particular DNA sequence and a characteristic related to human health allowed many researchers to argue that genetic data need not be considered sensitive in nature. More recent research has shown however that large portions of DNA contain many components that had a previously un-envisaged functional role in human development (and thus potentially health status). These include gene expression, silencing and the production of material that interacts with components in the cell in biochemical processes. In the last decade there has been a proliferation of research analysing specific environmental interactions with the genome in 'epigenetic studies' and more nuanced methods in understanding the dynamic interplay between genetic variation and environmental variation are being developed that build on the foundation of twin studies.[38] The growing body of understanding (of the functional role of various parts of the genome) means that assertions that genetic data used in research projects is not of a sensitive nature (i.e. do not provide an indication as to health status of the data subject) appear increasingly dubious. The increasing use of GWSs (and complimentary datasets) further weaken such assertions given that it is inevitable that a GWS will contain elements that can provide an indication of the health status of the data subject (even if such elements may themselves not be the focus of the particular research project).

With the advent of the GDPR, the need to show a potential link to health status has been removed.[39] Genetic data is now, by its own right automatically considered sensitive data.[40] The explicit description of genetic data as sensitive data can be viewed in two ways in terms of the practical difference it is likely to make for researchers. *First*, in some regards this step can be seen as a welcome clarification that genetic (personal) data is automatically sensitive, removing the nature to perform an assessment of whether this is indeed the case or not. Although more and more data is becoming and will become health data as time progresses (due to aspects such as increasing computing power, availability of complimentary data and complex analytic tools),[41] answering this question on a case-by-case basis would represent a consuming exercise that would have required both time and resources.[42] This is now dispensed with under the GDPR, given that when one is working with genetic data that is personal data (which one should often assume is the case) one is automatically required to treat it as sensitive data. Such an arrangement at least allows researchers to make arrangements for the consistent handling of personal data in research using genetic data. *Second*, whilst such certainty may be welcome from a planning perspective, it none the less confirms that researchers will always be subject to the more stringent requirements that apply to sensitive data when they are working with genetic data. This means that in most cases researchers will have to employ extra measures to ensure that the requirements pertaining to personal data are met. As subsequent sections of this paper will make clear these may be onerous and can relate to conditionality surrounding the use of consent as a legal base, the need to im-

33 Article 29 Working Party Document on Genetic Data (12178/03/EN WP 91)- 17 March 2004.

34 In theory genetic data could also be sensitive data under Directive 95/46/EC where it could be demonstrated that it also revealed data pertaining to the "racial or ethnic origin" of individuals (see Article 8(1)). This possibility will not be dealt with further in this paper however as it is beyond its scope.

35 Article 29 Data Protection Working Party, Advice Paper on Special Categories of Data ('Sensitive Data').

36 See annex to letter written by the Article 29 Working Party to the European Commission on Feburary 5th 2014 concerning the interpretation of health data, available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/otherdocument/files/2015/20150205_letter_art29wp_ec_health_data_after_plenary_annex_en.pdf.

37 Palazzo, A. & Gregory, T. 2014. The Case for Junk DNA. *PloS Genetics*, https://doi.org/10.1371/journal.pgen.1004351.

38 Amin, V., Behrman, J. R. & Spector, T. D. 2013. Does More Schooling Improve Health Outcomes and Health Related Behaviors? Evidence from U.K. Twins. *Economics of education review,* 35**,** 10.1016/j.econedurev.2013.04.004.

39 Chassang, G. 2017. The impact of the EU general data protection regulation on scientific research. *ecancermedicalscience,* 11**,** 709.

40 Hallinan, D., Friedewald, M. & De Hert, P. 2013. Genetic Data and the Data Protection Regulation: Anonymity, multiple subjects, sensitivity and a prohibitionary logic regarding genetic data? *Computer Law & Security Review,* 29**,** 317-329.

41 Malgieri, G. & Comandé, G. 2017. Sensitive-by-distance: quasi-health data in the algorithmic era. *Information & Communications Technology Law,* 26**,** 229-249.

42 One should however be aware that one of the potential obligations that will often be incumbent upon data controllers when they use sensitive (or special data) is to conduct an impact assessment (described in article 35 of the GDPR). See Section 6.

plement a data protection impact assessment (see Section 6), the creation of a Data Protection Officer[43] and other administrative requirements that pertain to sensitive data.[44]

5. The need to follow data processing principles

5.1. *Data processing principles cannot be consented away*

The data protection principles contained within the data protection framework are of crucial importance given that, in general, they must be adhered to in all cases of processing of personal data.[45] It is not possible for example for individuals to consent away the need to adhere to the data protection principles. Requirements such as accuracy, purpose limitation, data minimisation, storage limitation and privacy by design must thus be adhered to, even if consent has been obtained.

Whilst adhering to principles such as accuracy may be relatively straightforward (given the digital nature of genetic data when stored electronically), adhering to others may entail a difficult conceptual and logistic exercise. Researchers using big genetic data are likely to face a number of particular issues in addition to those faced by all forms of scientific research. As the discussion below indicates: not only may such adherence be difficult without reducing the value of a particular research experiment but, given the nature of big genetic data, it may even be difficult to understand what exactly compliance may entail. The authors hope that some clarity may be provided by the guidelines on the use of health data in medical research, which are still being prepared at the time of writing.[46] Such guidance is necessary because at present major uncertainties exist in terms of how data protection principles should be applied to fields such as computational genetics. The most problematic of these principles (from the perspective of computational genetics) are discussed below.

5.1.1. *Purpose limitation*

Purpose Limitation represents a requirement that data is not processed for purposes other than those that were the intended reason for collection.[47] Discerning the exact purpose of processing and the relevant boundaries that should apply to it in the context of complex computational genetics research may be problematic. This is because the very nature of the data mining operations that may be carried out may not always be amenable to simple and concise explanation. Not only are such operations complex but their goal may be vague such as looking for correlations between various sequences and physically observed phenomena. Researchers may not know what correlations they are looking for at the outset of the research. Given that researchers do not know exactly what correlations they are going to find (or even which ones they are searching for) or their potential significance in either research or privacy terms, it may strictly speaking be difficult to outline in a precise or concise manner what the exact purpose of the research is at its outset. This may also be made more difficult by the fact that data mining operations may be ‘opportunistic’ in nature. This is because the aim of such operations may be to discover unknown relationships and correlations between various genetic sequences and physical phenomenon.[48] In essence, such operations represent exercises at searching for ‘unknown unknowns’. The discovery of such relationships may in turn raise new questions and send the research (and the type of data processing operations it involves) in new directions. The nature of such research essentially means that the data mining operations that are being carried out (and their purpose) may be constantly changing because of new information that has been obtained. Given this, it may not be feasible to accurately and succinctly describe what the purpose of data collection is or adhere to such a description if it was indeed formulated. Formulating a purpose in a manner that was too precise or restrictive would likely severely restrict many types of research project in the area of computational genetics given that such research depends precisely on looking for previously unknown relationships that exist within human genomes and upon using such discoveries to drive further research.[49] Whilst as Section 8 discusses with regards to consent, the GDPR makes some accommodations in terms of the application of purpose limitation to matters of scientific research, much uncertainty remains.

5.1.2. *Data minimisation*

The principle of data minimisation is closely related to that of purpose limitation. It represents the notion that only the data that is required to meet the purpose that existed at the time of collection should be assembled.[50] The *raison d’être* of such a principle is to ensure that unnecessary personal data is not collected and thus reduce the risks to data subjects that they will suffer privacy harms. Such a principle in theory allows data controllers to process personal data for legitimate ends but prevents them from unnecessarily maximising the amount and types of data that is gathered. The GDPR itself explicitly confirms that this principle applies also to instances of processing for scientific research.[51]

In the world of big genetic data, however, the notion of data minimisation is often likely to be seen as being incompatible with the very nature of the research itself. This is because computational genetics (as with other forms of big data research) depends on maximisation (of often heteroge-

[43] Article 37 GDPR.

[44] Recital 26 of the GDPR makes clear the regulation does not apply to anonymized data.

[45] As Section 8 will discuss, where consent is not the legal base used, the application of some data processing principles may be limited.

[46] http://pr.euractiv.com/pr/gdpr-code-conduct-health-research-and-implications-fp9-159277.

[47] Ghani, N., Hamid, S. & Udzir, I. 2016. Big Data and Data Protection - Issues with Purpose Limitation Principle. *International Journal of Advances in Soft Computing and Its Application* 8**,** 116-121.

[48] Roche, M. & Berg, J. 2015. Incidental Findings with Genomic Testing: Implications for Genetic Counseling Practice. *Current Genetic Medicine Reports,* 3**,** 166-176.

[49] Takahashi, J., Pinto, L. & Vitaterna, M. 1994. Forward and reverse genetic approaches to behavior in the mouse. *Science,* 264**,** 1724-33.

[50] Article 1(c) GDPR.

[51] GDPR Recital 156.

neous forms) of data.[52] It is only through such maximisation that the requisite forms of complex data mining and analysis can be conducted. In the context of computational genetics itself, maximising such data means using the entire genome of one or more individuals (i.e. GWSs). The nature of such operations can perhaps be analogised as 'requiring the entire haystack to find the needle'. Although a meaningful finding may only represent a small sequence of DNA, it may nonetheless be necessary to use an entire GWS to carry out the research in question.

In order to find sequences of interest researchers often use approaches that will contrast test statistics of potential candidates with genome average scores.[53] This makes having a large amount of data to compare this signal to noise vital in adding validity to any potential findings. Although ultimately the overwhelming majority of genetic data in GWS data will not prove to be biologically relevant to a trait of interest, it is not known beforehand which portions of the genome this relates to, especially in candidate gene generation approaches.[54] In addition to this, evolutionary and population genetic analyses specifically aim to use neutrally evolving genetic markers across the genome (markers that do not have a role in disease) to make inferences about past relationships between present day populations.[55] The more markers available for study the more statistical power can be harnessed to make evolutionary inferences at higher resolutions. It is only with the accumulation of big data, in this case genome wide samples (and of many of them as possible) that such forms of analysis can be carried out. Restricting the size of the sample would reduce the power of the analytical techniques that could be used and the likelihood that any findings of interest could be secured. Given these factors, interpreting the concept of data minimisation is likely to prove challenging to researchers active in the field of computational genetics.

5.1.3. *Storage limitation*

Storage limitation is a central tenet of data protection. It essentially encapsulates the notion that data should not be stored longer than is necessary. Once the reason for processing the data in question no longer applies, data should be deleted. This reduces the risks of later improper use of the personal data by either the current controller or some other third party. In the context of scientific research, this would ideally mean that once the research was finished the data in question should be deleted. Whilst such a requirement may seem both logical and desirable from a data protection perspective there are two particular reasons why it may be difficult in areas such as computational genetics.

First, the fact that research projects in this area often do not have a clearly defined duration at the time the data was collected. Research may continue for longer than was initially planned because of further discoveries that occur after the data has been mined[56] (as subsection 5.1.2 above discusses) In addition to findings within current research projects new possibilities for analysis may be opened up by the subsequent availability of other complimentary data. This could be from data that has been made available online, from sources such as healthcare institutions or even from other research projects that have been concluded.[57] After enjoying initial results, projects may receive extra funding, extending both their scope and duration. This may often rely on using current research data for longer than was originally foreseen. Second is the need to fulfil what is often seen as an element of good practice in scientific research – making data sets available to subsequent researchers (discussed further in Section 3.3).[58] This may be in order to allow other researchers to verify their work or to allow researchers to use such data where their work builds on previous research. Therefore, researchers in the domain of computational genetics, as in other domains, are often placed under pressure to make their data available.

These needs, associated with research in general and computational genetics in particular, make adherence to the principle of storage limitation difficult. One possible solution to this conundrum might be to suggest that once initial findings have been made (say after search algorithms had identified certain sequences of interest), data relating to all other parts of the genome can simply be deleted. Given the size of the genome and the fact that researchers may in reality be interested in a tiny part or parts of it, this could mean that 99.9 percent of a data subject's genome sample could be deleted. Whilst this would no doubt be beneficial in terms of protecting a data subject's privacy and would go a long way towards respecting the requirement of storage limitation, it would not in reality be consistent with or permit many forms of research in computational genetics. Research projects often require continuous and ongoing processing of the original dataset. The search operations that are carried out may often be refined or even conceived of on an on-going basis. Initial discoveries and information may give rise to new questions and theories, events that will prompt researchers to look for new correlations and relationship within the data. In order for this to

[52] Hong, E. P. & Park, J. W. 2012. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics & Informatics,* 10**,** 117-122.

[53] Zhong, M., Zhang, Y., Lange, K. & Fan, R. 2011. A cross-population extended haplotype-based homozygosity score test to detect positive selection in genome-wide scans. *Stat Interface,* 4**,** 51-63, ibid.

[54] Although the physical manifestation of a trait may be well understood by clinicians and researchers its location and action in the genome is often unknown. Approaches including linkage mapping that show shared genomic segments with carriers of the disease allow identification of the genomic location of a trait of interest. See: Dawn Teare, M. & Barrett, J. H. 2005. Genetic linkage studies. *The Lancet,* 366**,** 1036-1044.

[55] Tischkoff, S. & Verreli, B. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annual Review Genomics Human Genetics,* 4**,** 293-340.

[56] Lyu, H., Huan, J., Zhimin, H. & Liu, B. 2018. Epigenetic mechanism of survivin dysregulation in human cancer.

[57] Clayton, E., Mizielinska, S., Edgar, J., Nielsen, T., Marshall, S., Norona, F., Robbins, M., Damirji, H., Holm, I., Johannsen, P., Nielsen, J., Asante, E., Collinge, J., Consortium, F. & Isaacs, A. 2015. Frontotemporal dementia caused by CHMP2B mutation is characterised by neuronal lysosomal storage pathology. *Acta Neuropathologica,* 130, 511-523.

[58] Mcguire, A. L., Hamilton, J. A., Lunstroth, R., Mccullough, L. B. & Goldman, A. 2008. DNA data sharing: research participants' perspectives. *Genet Med,* 10, 46-53.

occur, it will often be necessary to maintain and use the same GWS(s) that were used to produce initial results.

6. The need for an impact assessment

One of the novel requirements of the GDPR is the need perform a 'Data Protection Impact Assessment' (DPIA) in a number of circumstances where the proposed processing may "*represent a high risk to the rights and freedoms of natural persons*".[59] The GDPR does not exhaustively describe all the situations where a data protection impact assessment is required but does describe certain occasions where it shall be required, including situations that require "processing on a large scale of special categories of data". Given that genetic data is explicitly described by the GDPR as 'special data' (see discussion in Section 4) this raises the question of whether processing of the type described in paper (i.e. in areas of computational genetics) would fall within the description posed by Article 35(2) of the GDPR.

Whilst one cannot answer this question with certainty, it seems likely that scientific researchers involved in the processing of big genetic data will have to conduct a DPIA. Even though researchers may not always use the data of numerous individuals in a particular research project (it may only be a small group of people, a few individuals or even a single person),[60] the nature of big genetic data processing arguably means that such processing can be considered to be 'on a large scale'. This is because such data even if, from just a single individual in reality represents a vast amount of data given the size of the human genome.[61] A genome can provide an enormous amount of information relating to a diverse array of areas, from the likelihood of developing illnesses, to aspects linked to physical appearance, through to aspects linked to hereditary issues and ethnic origin.[62] In the future, the amount of information that will be capable of being deduced from big genetic data will only increase further. The possible for such future developments will have to be considered in any impact assessment. Furthermore, it is not only necessary to consider potential issues related directly to the data subject in question must, but also harms that may be created for other individuals that may share large portions of their genome (i.e. family and related individuals).

Even if, one was to consider that the concept of processing 'on a large scale' is not met in a particular instance, the Article 29 working party has identified other aspects that would strongly suggest that all instances of big genetic data processing are likely to require a DPIA. These include "the volume of data and the range of different data items being processed".[63] In addition, where sensitive data is combined with other datasets the chances that such processing represents a 'high risk to the rights and freedoms' of data subjects is likely to be greater. Furthermore, it states that another key factor to be considered is where "innovative use" of sensitive data is made or novel "technological or organizational" solutions are applied to it. Given that researchers often combine genetic data with other types of data (e.g. EHRs) and that the processes used are highly innovative it would be prudent to consider the processing of a few or even one GWS as being 'likely to result in a high risk' to rights and freedoms for the purposes of article 35 of the GDPR and therefore requiring a DPIA.

In terms of what exactly may be required concerning the form such an impact assessment should take or what substance it should have, there is at present much uncertainty. The authors of this paper would submit that this question is in need of further attention. Even a cursory consideration of the potential 'rights and freedoms' that could be theoretically at stake from the storage and processing of GWSs raises a potentially enormous range of issues. These range from harms on the individual level such as privacy harms to harms and questions related to access to healthcare to harms on a group or societal level (given many individuals may share genetic sequences of importance). A consideration of all such harms and the measures needed to mitigate them (if this is indeed what Article 35 is demanding)[64] would be a potentially enormous exercise demanding a truly multi-disciplinary perspective from disciplines such as ethics, law, genetics and sociology. How individual research groups are meant to either mobilise such resources or organize them (given the inevitable limited budgets that apply to most research projects) is a question that requires urgent attention.

[59] GDPR Article 35(1).

[60] Moltke, I., Albrechtsen, A., Hansen, T., Nielsen, F. & Nielsen, R. 2011. A method for detecting IBD regions simultaneously in multiple individuals–with applications to disease genetics. *Genome Research,* 21**,** 1168-80.

[61] The human genome is considered as the sequence of bases contained within the nucleus of cells of an individual. This entire genome is divided into 23 pairs of chromosomes of generally decreasing size from chromosome 1 to chromosome 22. The final pair of Chromosomes are the sex determining pair of Chromosomes, if an individual has two X chromosomes the resulting embryo will have female sex and if the individual has an X and a Y chromosome the embryo will have male sex. The sequence of the genome is what is used by researchers as it is the sequence of nitrogenous bases that varies between individuals and is of interest to researchers. The human genome contains roughly 3.2 billion pairs of nitrogenous bases, so roughly 6.4 billion bases in total. This is equivalent to 130 books that would require 95 years to read (https://www2.le.ac.uk/news/blog/2012/december/you-in-130-volumes-entire-human-genome-printed-for-exhibition).

[62] Many approaches for example use a simulations approach to model evolution under certain scenarios and compare to the actual data to estimate which scenarios would lead to the genetic variation presented in the study samples. An example of this approach uses 'the 'Markov Chain Monte Carlo' method which can generate billions of simulations and then present the simulation most likely to produce the experimental data. See: Konigsberg, L. & Frankenberg, S. 2013. Bayes in Biological Anthropology. *American Journal of Physical Anthropology,* 57**,** 153-84.

[63] Article 29 Working Party – Guidelines on Data Protection Impact Assessment (DIPA) and determining whether processing is likely to result in a high risk for the purposes of Regulation 2016/679 17/EN/WP248 Adopted 4 April 2007 (p9).

[64] For discussion on the potential breadth of article 35 see Kloza, D., Van Dijk, N., Gellert, R., Böröcz, I., Tanas, A., Mantovani, E., Quinn, P. (Brussels Laboratory for Data Protection & Privacy Impact Assessments (d.pia.lab)), Data protection impact assessments in the European Union: complementing the new legal framework towards a more robust protection of individuals - d.pia.lab Policy Brief No. 1/2017, 2017, ISSN 2565-9936.

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

Given that it is unlikely that most research outfits will have such expertise in-house there exist two obvious possibilities. The first is to hire external expertise in the form of consultancies or external advisors that specialize in such impact assessments. This is arguably not realistic given the limited resources available to most research groups. The second may be to create and utilize shared resources between research groups. This could involve the creation of a special unit within research institutions that would help and advise research groups on when and how they should perform a DPIA. Although this would involve expense for such institutions, it would inevitably be less than that needed were individual research groups required to secure such expertise separately.

7. The need to facilitate data subject rights

Data subject rights allow data subjects to ensure that their data is being processed both fairly and lawfully and, in a number of situations to exercise a level of autonomy over the processing of their personal data.[65] Data subject rights are closely linked to the need to respect data processing principles (discussed further in 5). Data controllers are required to facilitate such rights when processing personal data.[66] Whilst data protection rights may at first glance appear to entail simple operations concerning the processing of a data subject's personal data, the reality is that facilitation of such rights may often represent an onerous task for the data controller, particularly where the data processing involved or the organisational and structural arrangements around it are complex. Processing of GWSs for the purposes of computational genetics constitutes a form of processing that is particularly complex and thus creates problems in terms of the requirements incumbent upon data controllers. The need to facilitate data protection rights may also introduce complex procedural and administrative challenges given a particular data subject's data may be held along with the data of numerous data subjects as part of a large dataset. Such data is also likely to be pseudonymised and accessible to a potentially large amount of researchers that may be based at different research institutions.[67] Consequently, it may be necessary to establish formal administrative procedures in order to properly comply with data subject rights. For small research groups or institutions, or those using large and complex datasets such processes may entail a significant cost. As Sections 8 and 9 will discuss, such burdens are important to consider when choosing the legal base for processing given that the applicable data subject rights can vary according to whether consent or the research exception was opted for as the relevant legal base for processing. Some data subject rights that pose particular challenges to data subjects are discussed below.

7.1. *The right to information*

The right to information (RTI) is closely linked to the idea of fairness and transparency of process.[68] If data subjects are provided with the requisite information concerning their processing then they will be able to see that their data is being processed according to the law. The RTI is also closely linked to the ability to provide informed consent.[69] This is because consent is very much a two-way exercise that involves the flow of information between the data controller and data subject in both directions. On the one hand, the data subject signals that they understand how their data is to be used and the consequences of that processing. On the other, and in order to signal consent, it is necessary that the data subject receives the necessary information concerning the proposed processing in question. This is critical if he or she is truly to be able to give informed consent. Accordingly, the GDPR provides a number of informational requirements incumbent upon the data controller.[70] This includes inter alia a right to know the identity of all data controllers, the purposes of the intended processing, the legal base for processing,[71] the storage duration, any recipients of the data and information concerning transfer to any third country. In addition, it is necessary to transmit practical information on how data subjects can exercise their other data subject rights (some of which are described in the sections that follow below).

As is discussed in Section 8 on consent, fulfilling a number of these requirements will be problematic given the nature of computational genetics. This is particularly true concerning any description of the purpose of the processing, which for the reasons discussed in Section 5, will likely at best have to remain vague in nature. Were researchers to be too precise in delimitating a potential processing purpose it would likely make many normal processes in computational genetics impossible. As Section 8 discusses in more detail, the drafters of the GDPR were clearly aware of this problem and indicated that data subjects should be able to give a broader form of consent to "certain areas of scientific research"[72]

Such flexibility is not available in other areas. This includes the requirement to provide details of all controllers.[73] Given

---

65 De Hert, P. & Gutwirth, S. 2006. Privacy, Data Protection and Law Enforcement. Opacety of the Individual and Transparency of the Power. *In:* Claes, E., Duff, A. & Gutwirth, S. (eds.) *Privacy and the Criminal Law.* Antwerp - Oxford: Intersentia. Gutwirth, S., Leenes, R., De Hert, P. & Poulleto, Y. 2012. *European Data Protection: In Good Health?*, Springer.

66 Most data subject rights are contained within chapter 3 of the GDPR. Most make it explicitly clear that is the Data Controller that must facilitate such rights. This include the right to receive certain forms of information (articles 13 and 14), a right of access (article 15), a right of rectification (article 16), a right of erasure (article 17), a right to restrict processing (article 18) and a right to data portability (article 20).

67 The 1000 Genomes Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., Mccarthy, S., Mcvean, G. A. & Abecasis, G. R. 2015. A global reference for human genetic variation. *Nature,* 526, 68-74, ibid.

68 Lynskey, O. 2014. Deconstructing Data Protection: The 'Added-Value' of a Right to Data Protection in the EU Legal Order. *International and Comparative Law Quarterly,* 63, 569-597.

69 Mantovani, E. & Quinn, P. 2013. mHealth and data protection – the letter and the spirit of consent legal requirements. *International Review of Law, Computers & Technology*, http://dx.doi.org/10.1080/13600869.2013.801581.

70 GDPR Articles 13 and 14.

71 This matter is discussed further in Sections 7 and 8.

72 GDPR Article 33.

73 Article 13 GDPR.

the nature of collaboration between various researchers, groups and institutions in computational genetics, such a requirement may be no simple task.[74] This is especially true given that it may often be the case that new partners may join existing projects and gain access to the genetic data they are using after consent may have been obtained.[75] In the contemporary research environment, it is likely that a research consortium may involve many international partners, each operating in different legal jurisdiction in the EU and beyond.[76] Similar issues apply to question of 'storage duration' for the reasons elaborated in Section 5 (i.e. changing goals of a project, extensions to project duration and the need to publish research results).[77]

7.2. The right to access

The GDPR also allows data subjects at any time after collection to query data controllers on what personal data of theirs is stored and receive a copy of it in intelligible form. Such a right could be of importance where individuals want access to their data for family research purposes or where they may want to share their data with other scientific research projects. In order to facilitate this right of access it may be necessary to set up both administrative processes and data transfer infrastructure (so that where needed data can be transferred to data subjects). In big genetic data research involving potential multiple GWSs the question of a right of access is complicated for two reasons. The first is that data may be pseudonymised to increase both data subject privacy and security. This may make discerning the identity of the exact data subject difficult. Whilst discerning the identity of the data subject may be possible, it may be complex in terms of the technical and administrative steps required and thus be onerous for researchers. The second complication is that data generated within research projects may be aggregate in nature and thus may not relate to a specific individual as such. Conclusions are more likely to be in the form of (potentially very) low level correlations between various DNA sequence variations that may have limited relevance to specific individuals and may not even be considered as personal data. As a result, access requests may be unlikely to provide more than raw sequence data, which may be of limited use to data subjects.

7.3. A Right of right to erasure

The GDPR also foresees a right of erasure (commonly known as a 'right to be forgotten').[78] This allows individuals to demand that data controllers delete their personal data if one of a number of conditions are met. These include that the purposes for collection in the first place no longer exist or instances where individuals explicitly request the data to be deleted.[79] Where individuals who have given their consent for their data to be used in research directly request that research data pertaining to them is deleted however, researchers will have to comply with such requests.[80] Complying with such requests however may be problematic for a number of reasons. This includes (as Section 3.3 discussed) the increasingly common practice of publishing research data, or at least making data available to other researchers on request. Not doing so (for example as a result of erasure requests on the part of data subjects) may be seen as reducing the potential reliability of the results in question. Another issue is related to the ever growing understanding of the potential importance of certain sequences in genomic research projects and the direction of such research projects. As Section 5 discussed, such projects may start with only vague goals or strategies. This may involve applying new techniques to existing data. Erasure requests (where they occur) may make this problematic however and may harm the ability to continue or start research projects where it is necessary to continue using the same data set.[81] This can be envisaged in research projects that utilise multi-stage 'pipelines' where primary genetic data is processed in several stages in order to obtain results. Different individuals at different institutions often carry out the separate stages of the pipeline. Removal of an individual sample when an investigation has reached this stage may require restarting the whole pipeline from the beginning.

[74] See for example: Malaspinas, A. S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., Bergstrom, A., Athanasiadis, G., Cheng, J. Y., Crawford, J. E., Heupink, T. H., Macholdt, E., Peischl, S., Rasmussen, S., Schiffels, S., Subramanian, S., Wright, J. L., Albrechtsen, A., Barbieri, C., Dupanloup, I., Eriksson, A., Margaryan, A., Moltke, I., Pugach, I., Korneliussen, T. S., Levkivskyi, I. P., Moreno-Mayar, J. V., Ni, S., Racimo, F., Sikora, M., Xue, Y., Aghakhanian, F. A., Brucato, N., Brunak, S., Campos, P. F., Clark, W., Ellingvag, S., Fourmile, G., Gerbault, P., Injie, D., Koki, G., Leavesley, M., Logan, B., Lynch, A., Matisoo-Smith, E. A., Mcallister, P. J., Mentzer, A. J., Metspalu, M., Migliano, A. B., Murgha, L., Phipps, M. E., Pomat, W., Reynolds, D., Ricaut, F. X., Siba, P., Thomas, M. G., Wales, T., Wall, C. M., Oppenheimer, S. J., Tyler-Smith, C., Durbin, R., Dortch, J., Manica, A., Schierup, M. H., Foley, R. A., Lahr, M. M., Bowern, C., Wall, J. D., Mailund, T., Stoneking, M., Nielsen, R., Sandhu, M. S., Excoffier, L., Lambert, D. M. & Willerslev, E. 2016. A genomic history of Aboriginal Australia. *Nature,* 538, 207-214. Data was collected from several sources including malarial disease based studies and biological anthropological research. After data was collated analysis was performed in institutions in three different continents and results compiled when the paper was drafted.

[75] For a discussion on the large amount of potential data controllers in a big genetic data research project and potential methods that can be deployed to increase privacy see: Al Aziz, M. M., Hasan, M. Z., Mohammed, N. & Alhadidi, D. 2016. Secure and Efficient Multiparty Computation on Genomic Data. 278-283.

[76] Mead, S., Uphill, J., Beck, J., Poulter, M., Campbell, T., Lowe, J., Adamson, G., Hummerich, H., Klopp, N., Rückert, I.-M., Wichmann, H. E., Azazi, D., Plagnol, V., Pako, W. H., Whitfield, J., Alpers, M. P., Whittaker, J., Balding, D. J., Zerr, I., Kretzschmar, H. & Collinge, J. 2012. Genome-wide association study in multiple human prion diseases suggests genetic risk factors additional to PRNP. *Human Molecular Genetics,* 21, 1897-1906.

[77] See for example Consortium, T. U. K. 2015. The UK10K project identifies rare variants in health and disease. *Nature,* 526, 82-90.

[78] GDPR Article 17.

[79] Articles 17(a) and (b) respectively.

[80] As Section 9 discusses further, where scientific processing is the selected base the applicability of the 'right to be forgotten' can be limited meaning that in certain contexts scientific researchers may not have to honour such requests. See GDPR Article 17(1)(f).

[81] Andrews, L. 1994. *Social, Legal, and Ethical Implications of Genetic Testing,* Wshington (DC), National Academic Press.

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

8. Consent as 'Default' legal base?

8.1. Researchers have a choice of legal base

A *sine qua non* for the processing of personal data is the existence of a legal basis for processing given its context and purpose. As with its predecessor, the GDPR sets out a (expanded) number of potential legal bases that can be used to justify the processing of personal data.[82] The choice of legal base is important for researchers not only because it is necessary to find one that is correct given the particular context involved, but also because each legal base comes with both its own conditionality and opportunities that will be relevant for the research project concerned. The choice of legal base may also importantly influence the number and extent of data protection rights which must be complied with by data controllers (including inter alia the 'right of erasure' – see Section 7.3). It will also influence the flexibility researchers have in terms of the applicability of the principle of 'purpose limitation' . Choosing the correct legal base for the processing of data in research is thus of paramount importance. For computational genetics research, only two of the legal bases outlined in the GDPR are likely to be of relevance. These are either that the 'explicit consent' [83] of the data subject has been obtained or that such processing is necessary for" scientific or historical research purposes or statistical purposes" in line with Member State law.[84]

Of these two options, explicit consent is often considered the 'default option' for scientific researchers.[85] This is for a number of reasons. In the modern research context, it is often seen as being the most appealing from an ethical perspective.[86] From a legal perspective, it would appear to be privileged by the GDPR, which only permits the use of the scientific research exception when such use is "necessary".[87] The inclusion of this requirement clearly indicates that this option should only be considered when the use of consent as a legal basis is clearly not suitable.[88] In contrast, the absence of the word 'necessary' in the description of explicit consent as a legal base is notable, indicating its potential use and seeming default status as a legal base in most instances (including scientific research).[89]

8.2. Key problems surrounding consent

Whilst consent may clearly be considered the general default for the processing of data in the context of scientific research, the authors of this paper would argue that in many instances it is nonetheless unsuitable, especially in instances where many GWSs or large amounts of complimentary data (e.g. health records are used in conjunction with the primary genetic data). In particular, the three core requirements of such consent taken together i.e. that it be explicit, specific and informed would seem to be problematic.[90] Each of these elements presents particular problems linked to the nature of research in areas such as computational genetics. For consent to be both explicit and specific for instance, it must be informed. How potential data subjects can be precisely informed about what will happen with the data in the context of many computational genetics research programmes is difficult to imagine given the issues outlined in Section 5.[91] Given this, such consents are likely to be necessarily vague and broad in nature. Fortunately, for those conducting research in this and other areas the GDPR appears to recognise such problems and seems to call for the special context of scientific research to be recognised. Recital 33 states:

> *It is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research. Data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose.*

The contents of this recital are noteworthy given that in general the GDPR further expands (see Section 7) and even explicitly outlines the informational requirements that apply to data controllers (in comparison with Directive 95/46/EC). Rather than requiring a clear description of the purpose for processing as is normally the case, the GDPR appears to allow individuals to consent simply to research in 'certain areas'. This appears to open up the possibility for broad forms of consent that could for example include 'research into genetic origins of disease' or 'research on aspects of historical human

[82] Articles 6 and 9 (relating to special data) of the GDPR contain a much expanded potential list of legal bases that will allow the processing of personal and even sensitive personal data. Some of these grounds represent clarifications and precisions of several grounds that would have fallen within a single ground under Directive 94/46/EC. This includes a number of grounds in the GDPR (including for scientific research in article 9(2)(j) that were understood to be implicitly falling within the grounds of public interest in Directive 95/46/EC.

[83] GDPR Article 9(2)(a)

[84] GDPR Article (9(2)(a).

[85] Quinn, P. 2017. The Anonymisation of Research Data — A Pyric Victory for Privacy that Should Not Be Pushed Too Hard by the EU Data Protection Framework? *European Journal of Health Law,* 24, doi 10.1163/15718093-12341416.

[86] Carter, P., Laurie, G. & Dixon Woods, M. 2015. The social licence for research: why care.data ran into trouble. *Journal of Medical Ethics,* doi:10.1136/medethics-2014-102374.

[87] Article 9(2)(j).

[88] Article 29 Working Party Document on the processing of personal data relating to health in electronic health records (EHR) (2007) WP 131

[89] This is currently the position in many national laws on data protection. See for example article 23 of the Dutch Data Protection Act (Wet bescherming persoonsgegevens).

[90] For a good (pre-GDPR) overview see the Article 29 Working Party Opinon (15/2011) on Consent. GDPR Article 13 describes information that must be provided by a Data Controller to the Data Subject when the data in question has been directly obtained from the Data Subject. Article 14 describes information that must be provided to the data subject when the data in question has not been provided directly from the data subject.

[91] See also: Kaye, J. 2012. The Tension Between Data Sharing and the Protection of Privacy in Genomics Research. *Annual Review of Genomics and Human Genetics,* 13, 415–431.

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

migration' through genetic analysis.[92] It is important to remember however that the discretion seemingly being hinted at towards researchers is located within a recital and thus not directly legally binding. Caution in not over interpreting its meaning is therefore necessary. The general requirements of the GDPR concerning consent for sensitive data still apply and retain the force of hard law. In addition, recital 33 itself also confirms that the principal of purpose limitation still applies, confirming that the intention of this recital is by no means to grant *carte blanche* to researchers. Whilst it is wise therefore to assume the obligations upon data controllers to ensure that consent for the processing of sensitive data is "explicit, specific and informed" still applies to scientific researchers, such obligations should however be read in the special context that is scientific research.[93] This will entail finding and pursing a delicate balance that will in each case need to be judged according to the specific context involved.

8.3. *Examples of practical problems surrounding consent*

At present, most large genetic research projects usually rely on participants providing 'explicit consent' as the legal basis for the processing of genetic personal data.[94] Consent is not only seen as the standard legal requirement but also the most ethically acceptable option.[95] Opting for consent may thus not only serve a legal purpose (i.e. related to compliance with laws on data protection and privacy) but also increase the chances that ethical approval will be secured. How researchers go about obtaining consent has changed and developed throughout the years with large projects now often going beyond simply providing a consent form and providing materials with project information provided by researchers in order to help participants understand the research in question.[96] In better examples, researchers will explain which legal rights participants are entitled to, including those found in the data protection framework. This includes rights to access data, to be forgotten, and obligations such as data minimisation and storage limitation. In line with the seeming breadth described in article 33 of the GDPR, (including in the example below) researchers will also provide a broad (but informative) description of the goals of their research and the general methods used.

Practical Example – 100k Genome Project[97]

The 100,000 genomes project (100kGP) is an example of a project that highlights the direction genomic research is heading in the future. 100kGP is a multi-billion pound undertaking by Genomics England, a company wholly owned by Healthcare England. The project aims to harness the vast resources at the disposal of the UK National Health Service (NHS) in terms of patient numbers, patient data and patient records to generate a resource for clinicians to deliver a more personalised 'precision medicine' and to give researchers a statistically powerful array of data that will allow advancement in the understanding of development of diseases and potential therapeutic options. In addition to the personal genetic data from patients, organisers hope to continually update the database with 'long term patient health and personal information' to aid in the analysis to be undertaken by researchers. The project is mostly focused on the study of cancers and rare diseases (including rare forms of cancer). The project states in the first of its four major aims to 'create an ethical and transparent programme based on consent'.[98] Consent appears to be the underpinning of the measures taken to reassure patients regarding sensitive and ethical handling of their 'personal' genetic data and to comply with the law.[99]

Consent Process - Prior to reading and signing the consent form potential participants are provided with an information sheet containing explanations of the background science of the project as well as the measures to be taken to ensure that privacy and anonymity is maintained as well as possible. There are variations of these forms depending whether they are being provided for adults, parents or relatives of deceased individuals. There are also variations depending whether a participant is a cancer patient, patient with a rare disease or an infectious disease. Such information can include explanations on the nature of DNA itself, a general project outline, descriptions of the perceived risks of participating, information concerning the extraction and storage of samples, policies concerning access to data for researchers and organisations, security of data, results of research, additional findings, carrier testing and implications for relatives/descendants and information concerning how to withdraw from the project. In addition to the information sheet, there is additional information on the 100kGP web page, including documents to download and read, web pages with questions and answers, and video testimonies from participants, clinicians, researchers and administrators of the project. Information is split into areas including the vision of the project and its history, and possible concerns that may arise for patients. Whilst it is possible to make some criticisms, the authors of this paper believe that such efforts can be regarded as good example of genuine intention to meet the requirements of consent for sensitive data. Given the increased room for manoeuvre seemingly outlined in recital 33 of the GDPR the provision of such information is arguably sufficient to meet the requirements for informed consent to occur in instances of scientific research.[100]

92 The 1000 Genomes Project, which is an example of a study where participants were informed that their data would be used not only for population genetic history studies but also for studies into the evolution of disease related variants.

93 Shabani, M. & Borry, P. 2017. Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation. *Eur J Hum Genet*.

94 Mcguire, A. L. & Beskow, L. M. 2010. Informed Consent in Genomics and Gentic Research. *Annual review of genomics and human genetics*, 11**,** 361-381.

95 Friedewald, M. & Hallinan, D. 2015. Open consent, biobanking and data protection law: can open consent be 'informed' under the forthcoming data protection regulation? *Life Scienes and Social Policy*, 11**,** 1-36.

96 Mcguire, A. L. & Beskow, L. M. 2010. Informed Consent in Genomics and Gentic Research. *Annual review of genomics and human genetics*, 11**,** 361-381.

97 *The 100,000 Genomes Project Protocol v3, Genomics England. doi: 10.6084/m9.figshare.4530893.v2. 2017.*

98 From the outset, organisers of the project have stated that access to the data will be limited and security protocols updated and modified as the project progresses. See *Genomics England Website* - https://www.genomicsengland.co.uk/the-100000-genomes-project/.

99 Other measures include the creation of a specific ethics and advisory board that will scrutinise how patients receive results and

9. The use of the 'Scientific research exception'

9.1. An alternative to consent as legal basis

In addition to 'explicit consent', another potentially relevant legal base is where such processing may be in the "public interest".[101] This provision has thus far been used by Member States in their transposition of Directive 95/46/EC (and in other legislation) to permit processing of sensitive data for a range of purposes, including for scientific research.[102] Within Article 9(2) of the GDPR, this exception has been split into a number of specific grounds that were understood as being implicitly included within article 8(4) of Directive 95/46/EC. Of most relevance to this paper is article 9(2)(j) of the GDPR which states that processing of sensitive data can occur where such processing

> *is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.*

As with the GDPR, consent was seemingly favored as the exception of preference under Directive 95/36/EC where the 'public interest' option was, according to the Article 29 working party, to be read narrowly.[103] Use of this base for the processing of data could only be made where it was both necessary and proportional. This order of preference of potential legal bases was logical given that obtaining consent was not only preferable in an ethical sense, but also less problematic given the types of research and the data that was used at the time the directive was created. The ability to collect and mine large sets of useful datasets was limited in comparison to what is possible today.[104] Researchers were therefore limited to relatively small datasets for which obtaining consent was in most cases arguably feasible. This situation has gradually been changing however, especially given the increased amount of scientific research that relies upon big data analytics and the re-use and sharing of research data.[105] The GDPR (and the protracted discussions concerning its formulation) appear to recognize this, calling for a broad interpretation of the concept of scientific research that applies not only to research by public bodies such as universities but also private and commercial entities also.[106]

This exception, although capable of interpretation in a broad manner nonetheless comes with important conditionality.[107] Measures that seek to utilise this exception must fall within circumstances clearly described within Member State law and be both necessary and proportional in order to achieve the public interest related aim in question (i.e. consent should not be suitable).[108] Meeting such requirements may sometimes require conditionality that is demanding and difficult to fulfil. Such conditionality is demanded in general terms by the GDPR and is to be specified in more specific terms by the relevant Member State legislation in place in a certain jurisdiction. There will therefore still be a great deal of heterogeneity in such law across the EU even despite the choice of a regulation (which is usually synonymous with harmonization).[109] It may for example require the introduction of complex administrative measures to ensure security, the use of encryption or the pseudonymisation of research data to a high degree.

consent policies. The project has also undertaken engagement and outreach events with patients, clinicians and researchers.

100 https://www.theguardian.com/science/political-science/2015/mar/10/privacy-and-the-100000-genome-project There have been concerns raised by journalists and members of the public regarding privacy concerns of 100kGP. With critics focusing on the impossibility of obtaining true informed consent and the lack of transparency over the possibility of third parties deanonymising individuals and violating their privacy. In a newspaper investigative piece the authors allege that Genomics England are wilfully not disclosing the fact that individuals data is psuedonymised and not anonymous as stated on the web site. They go on to state that GE is trying to portray the two terms as synonyms when in fact they are very different and with great implications for participants in the project.

101 This possibility was described in Article 8(4) of Directive 95/46/EC.

102 In Germany, for example Article 4(1) of the data protection act permits the use of health data without consent for scientific research. Article 13 states that health data "may be collected where the scientific interest in carrying out the research project substantially outweighs the data subject's interest in excluding collection and the purpose of the research cannot be achieved in any other way or would otherwise necessitate disproportionate effort" For more discussion of this and similar provisions in other European jurisdictions see: Deliverable 9.3 'ELSI tools for standardization and harmonisation to use data from different biobanks' from the project BIOSHARE, Grant No. 261433.

103 Article 29 Working Party Opinion on Working Document on the processing of personal data relating to health in electronic health records (EHR) 00323/07/EN WP 131 Adopted on 15 February 2007 p8.

104 Kohane, I. 2011. Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics*, 12, 417 - 428.

105 Quinn, P. 2017. The Anonymisation of Research Data — A Pyric Victory for Privacy that Should Not Be Pushed Too Hard by the EU Data Protection Framework? *European Journal of Health Law*, 24, doi 10.1163/15718093-12341416.

106 GDPR Recital 159.

107 This was also the case with Directive 95/46/EC with recital 34 stating "Whereas Member States must also be authorized, when justified by grounds of important public interest, to derogate from the prohibition on processing sensitive categories of data where important reasons of public interest so justify in areas such … scientific research … it is incumbent on them, however, to provide specific and suitable safeguards so as to protect the fundamental rights and the privacy of individuals".

108 The requirements of necessity and proportionality stem from the requirements the European Court of Human Rights has laid down under Article 8 of the European Convention on Human Rights for measures that may infringe upon the private life of individuals. For more discussion on this see Article 29 Working Party Opinion on the processing of personal data relating to health in electronic health records (EHR) (2007) WP 131.

109 Article 9(4) of the GDPR furthermore states: "Member States may maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health."

The conditionality that is required means that the use of the research exception cannot be considered as constraint free. Imagine for instance conditionality that requires an extremely high level of pseudonymisation.[110] Where such demands are particularly high they may serve to lessen the potential uses of the data in question.

9.2. *Advantages of using the scientific research exception*

The most obvious advantage of using the scientific research exception for processing is that consent does not have to be obtained. This may be important in many forms of research (including computational genomics) where gaining the consent of all data subjects would be problematic.[111] This includes where data subjects may no longer be alive, where the data subjects are minors or do not have capacity to give consent and where data has been taken from other sources (i.e. previous experiments or health records), thus making locating the data subjects in question difficult. The existence of another potential legal base provides a further option where the need to secure consent would be a major hindrance to a particular research project.

In addition to the obvious advantage of not having to obtain consent from data subjects, the 'scientific research' option has a number of potential advantages that would seemingly be suited to domains such as computational genetics. Some of these advantages are outlined directly in the GDPR itself, whilst in other cases the regulation leaves the possibility open for Member States to curtail a number of data protection rights in their respective legislation.

In terms of the first category, the GDPR limits the application of informational rights (discussed in Section 7) in instances where the legal basis for processing is 'scientific research'.[112] This limitation is noticeable in two regards. Firstly, it applies only in terms of the informational rights outlined in article 14 of the GDPR and not Article 13. Though appearing broadly similar, the former only applies in instances where the personal data concerned has not been collected directly from the data subject (unlike article 13 which concerns instances where data has been collected directly from the data subject). This is consistent with the idea that such an exception should not apply to instances where processing is based upon the consent of the individual given that often in such cases personal data is taken directly from the individual in question. Second this exception only applies in instances where complying with information rights would "involve a disproportionate effort … in so far as the obligation referred to in paragraph 1 of this Article is likely to render impossible or seriously impair the achievement of the objectives of that processing". The presence of this requirement means that informational rights described in the GDPR cannot simply be ignored in the cases of scientific research. In place of such a blanket assumption, it is necessary to look at each one and discern whether, given the particular context, they are proportional or not. The authors of this paper would argue that given the elements discussed in Section 5 (i.e. concerning purpose limitation in matters of computational genetics)[113] it may indeed be acceptable to argue that compliance with a number of the informational requirements would, in certain contexts, be disproportionate and likely to impair the purposes of the processing (i.e. the research aim in question). This includes particularly the intended purpose, precise descriptions of all forms of processing and storage duration.

In terms of the second category described above, the GDPR allows Member States, through national legislation to limit data subject rights where the processing basis is 'scientific research'. In particular, the GDPR states:

> *Where personal data are processed for scientific or historical research purposes or statistical purposes, Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18 and 21 subject to … conditions and safeguards*[114]

The rights referred to relate respectfully to the 'right of access', a 'right to rectification', a 'right to restrict the processing' and a 'right to object to automated decisions that concern him or her'. The first may be relevant, for example, where individuals want a copy of their genome in order to provide it for third party services or scientific research. Given that access to a particular individual's genetic sample may be difficult in both administrative and technical terms (discussed in Section 7); complying with such a request may arguably in some instances be considered as cumbersome. This may especially be the case where there are numerous data subjects and the data has been heavily pseudonomysed. In such instances, where Member State law permits it, researchers may not have to comply fully with such requests when data has been processed under the scientific research exception. Where such exceptions do apply, they will create valuable room for researchers in instances where application of all of the usual data subject rights in a particular research context would have proved burdensome. In the UK for instance the recent Data Protection Bill,[115] which is currently going through parliament

[110] Indeed the GDPR itself seems to demand this in Article 89(1) stating: "Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, … Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner."

[111] Boddington, P., Curren, L., Kaye, J., Kanellopoulou, N., Melham, K., Gowansa, H. & Hawkinsa, N. 2011. Consent Forms in Genomics: The Difference between Law and Practice. *The European Journal of Health Law,* 18, 491-519.

[112] GDPR Article 14(5).

[113] See also Boddington, P., Curren, L., Kaye, J., Kanellopoulou, N., Melham, K., Gowansa, H. & Hawkinsa, N. 2011. Consent Forms in Genomics: The Difference between Law and Practice. *The European Journal of Health Law,* 18, 491-519. and Kaye, J. 2012. The Tension Between Data Sharing and the Protection of Privacy in Genomics Research. *Annual Review of Genomics and Human Genetics,* 13, 415–431.

[114] GDPR Article 89(2).

[115] The derogations are dealt with in Section 6 of the bill. A copy of the draft bill was available at the time of writing at https://publications.parliament.uk/pa/bills/cbill/2017-2019/0190/18190.pdf A description of the various derogations foreseen in UK law is available from the Open Rights Group at: https://www.openrightsgroup.org/assets/files/pdfs/dcms/Summary%20of%20GDPR%20derogations%20in%20Data%20Protection%20Bill.pdf.

has foreseen derogations for processing by researchers for "an individual's rights to access (Art. 15), rectify (Art.16), restrict further processing (Art. 18) and object to processing (Art. 21) where this would seriously impede their ability to complete their work", and providing that appropriate organisational safeguards are in place to keep the data secure. In other jurisdictions, such options will only be available to researchers where they specifically exist in the relevant Member State law (being applicable in the particular context in question and where the use of consent is unsuitable.

Practical Example – The Social Care Information Centre/ NHS Digital

The Social Care Information Centre (SCIC) was set up in the UK by the Health and Social Care Act 2014 (previously the Health and Social Care Act (2008)). It has recently been renamed 'NHS Digital'. This entity has many aims, including the facilitation of research that can help solve health problems. This includes genomic research. The SCIC has access to a wealth of high quality data (e.g. patient health records) available to the NHS. Although it is legally able to share such data with researchers where it is capable of bringing about "the promotion of health", in practice it provides for an 'opt out procedure' whereby individuals can signal that they do not want to be involved (such an option should not be confused with explicit consent which is more synonymous with an 'opt in').[116] In doing so, it uses the possibility that exists under Directive 95/46EC for member states to create laws allowing for the processing of sensitive personal data where it is in the public interest. In order to receive data from SCIC/NHS Digital, potential research institutions must demonstrate that they will store and process data safely and in line with the law. They must also sign a contract to that effect. Where possible data will be heavily pseudonymised before transfer.[117] In most cases, research proposals to use data from patient records will have to undergo ethical review (discussed further in Section 10 below).

The availability of such an option for the provision of research data without the explicit consent of the data subject does not however mean that it is frequently used. SCIC has only approved the transfer in restricted cases and only where the use of consent is clearly unsuitable. Even where potential data transfers have been approved by SCIC/NHS Digital, other obstacles can still exist to data transfers. This occurred with the case of the of the controversial 'care.data' platform in which SCIC/NHS Digital was to make patient data available for research. Various organisations within the NHS refused to transfer patient data to SCIC because of ethical concerns over the lack of explicit consent on the part of patients even though such transfers would have been legally sound (this issue of reticence is discussed further below).[118] Following much controversy care.data. was eventually scrapped.

10. The critical role of ethics bodies

Despite the clear existence of a legal ground for the processing of sensitive data for research purposes that does not require consent, regulatory authorities and ethics bodies have, in many cases, been reticent to use this option, preferring to insist that researchers obtain consent or use anonymised data.[119] This has been most noticeable amongst ethics bodies, which, often to the consternation of research scientists, have become ever more demanding in the requirements that they formulate when personal data is to be used. In many instances, such requirements go beyond what is required by both European and national data protection law.[120] This situation is, in the opinion of the authors of this paper, likely to continue after the GDPR comes into force in May 2018 even though the GDPR explicitly foresees scientific research as a legal basis for the processing of sensitive data.

The position that ethics bodies or other similar entities take can be critical because depending on national, local, or sectorial regulations, they may have the ultimate say in deciding whether a research proposal is approved or not.[121] In particular, many ethics bodies seem to have developed an aversion for research that aims to utilise personal data, which is not accompanied by the consent of the data subjects involved. Such reticence exists even though the law (through the public interest option) envisages the possibility for the use of personal health data without consent. This has, according to some research scientists, led to situations where research that is legally permissible is not even contemplated because of an over zealous approach by ethics committees.[122]

The reluctance of regulators and ethics bodies to grant permission for the use of the health data without consent often means that researchers are faced with a *de facto* choice between gaining consent or anonymisation. In reality, however either of these options may be difficult or even unachievable in many research contexts (see Section 3.3). Major problems may exist because data that is truly anonymous may often offer little or no potential in terms of research value.[123] As the authors discussed in Section 3.3, speaking of a GWS as

[116] For a review of these procedures see: *National Data Guardian for Health and Care Review of Data Security, Consent and Opt-Outs also known as the Caldicott review. Available at:* https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/535024/data-security-review.PDF.

[117] Guidelines for interested researchers can be found at https://www.mrc.ac.uk/documents/pdf/obtaining-data-from-nhs-digital-v101016/.

[118] Boddington, P., Curren, L., Kaye, J., Kanellopoulou, N., Melham, K., Gowansa, H. & Hawkinsa, N. 2011. Consent Forms in Genomics: The Difference between Law and Practice. *The European Journal of Health Law,* 18, 491-519.

[119] Quinn, P. 2017. The Anonymisation of Research Data — A Pyric Victory for Privacy that Should Not Be Pushed Too Hard by the EU Data Protection Framework? *European Journal of Health Law,* 24**,** doi 10.1163/15718093-12341416. Carter, P., Laurie, G. & Dixon Woods, M. 2015. The social licence for research: why care.data ran into trouble. *Journal of Medical Ethics,* doi:10.1136/medethics-2014-102374.

[120] Boddington, P., Curren, L., Kaye, J., Kanellopoulou, N., Melham, K., Gowansa, H. & Hawkinsa, N. 2011. Consent Forms in Genomics: The Difference between Law and Practice. *The European Journal of Health Law,* 18**,** 491-519.p5.

[121] Kaye, J. 2012. The Tension Between Data Sharing and the Protection of Privacy in Genomics Research. *Annual Review of Genomics and Human Genetics,* 13**,** 415–431.

[122] Carter, P., Laurie, G. & Dixon Woods, M. 2015. The social licence for research: why care.data ran into trouble. *Journal of Medical Ethics,* doi:10.1136/medethics-2014-102374.

[123] Fears, R., Brand, H., Frackowiak, R., Pastoret, P., Souhami, R. & Thompson, B. 2014. Data protection regulaiton and the promotion

anonymous may in reality be a misnomer. Even if this were not the case, genetic data is far more useful for research purposes where it contains personal (or quasi-personal identifiers) that allow the data in question to be analyzed within specific contexts. These could include a range of useful metadata that link genetic data to a particular health record, or attributes that allow a number of records to be linked together (e.g. showing familial relationships).[124] The dichotomy of 'consent or anonymise' presented by many ethics bodies may in the context of big genetic data research projects thus represent an unappealing choice for researchers. Given the potential for harms to research, the authors of this paper would call for further efforts to be made to highlight the problems of such a dichotomy, in particular by European and national data protection regulators and ethical bodies.

11. Conclusion

Computational genetics is undergoing a revolution. A number of developments have fuelled this revolution. Chief amongst these is the increasing ability to produce (rapidly and for low cost) GWSs. These can be mined repeatedly because of increases in computing power. The possibility to access and share various forms of potentially compatible information throughout the online-connected world have not only allowed for more research opportunities but also changed the way we view genetic data in legal terms. In particular, it has become increasingly difficult to regard any large sequence of DNA (let alone a GWS) as being anonymous. This is because the increase in both computing power and processing algorithms taken together with the online availability of enormous amounts of complimentary data mean that it is becoming ever more likely that such samples can be linked to identifiable individuals. Therefore, it is generally accepted that meaningful sequences of DNA should be considered as personal data.

The consequences of this change in the legal perception of genetic data is important because it means that one should assume that the EU's data protection framework applies to the processing of genetic information including the mining of GWSs. From May 2018, the bedrock of this framework will be the General Data Protection Regulation. This regulation explicitly describes genetic data as 'special' (formerly known as 'sensitive') data. This means that where personal data is genetic in nature, further requirements will be incumbent upon those processing genetic data. These include requirements pertaining to data processing principles (i.e. that apply to processing of personal data in general), the need to facilitate the data protection rights of data subjects and the need to ensure that there is a correct legal base for the processing of the research data in question. Researchers may also have to perform a complex and demanding data protection impact assessment.

As Section 5 discussed, the very nature of computational genetics means some of these requirements may not only be difficult to envisage but may even be seemingly contradictory with the very nature and goals of computational genetics. Whilst the GDPR may go some way in easing concerns about the compatibility of the notion of purpose limitation with various forms of scientific research, problems remain, especially in terms of the concepts of data minimisation and storage limitation. Given that computational genetics increasingly relies on the 'opportunistic mining' of GWSs, the idea of data minimisation is difficult to square with research techniques that require the use of the entire genome (and often large complimentary datasets e.g. healthcare records) to function. Similar problems are likely to apply when considering the issue of 'storage limitation'. In many forms of computational genetics research, the exact goal may not be well defined at the project's outset. Continuous new discoveries may not only change a project's direction but also its duration on an ongoing basis. Other problems may be created by the need to comply with data subject rights. This includes for the example the need to be provided with a range of information concerning the project. Again, identifying the precise duration of the project in, and all types of processing that may occur may be difficult or impossible at the outset of the project. Outlining the duration of the project will also provide further difficulties. Further difficulties may exist where researchers are required to honour requests by research subjects to halt the processing of data or to erase it, especially where such data may be central to any findings that have occurred within a project.

In terms of selecting a correct legal base for processing, the GDPR provides scientific researchers with two main options. The first is to obtain explicit consent from all data subjects. Such consent must accordingly be explicit, specific and informed. Ensuring these conditions are met in the case of computational genetics projects is no easy task given the open and opportunistic nature of the data mining processes that occur in such projects. The GDPR, in its recitals, however goes someway to recognising such difficulties in stating that data subjects should be able to give consent to research projects where the intended purposes of processing is not narrowly defined. Such a recognition does not obviate the need to obtain explicit consent but does create some extra manoeuvring space for computational geneticists in terms of the information they must provide to accompany such consent. Researchers will have to tread a careful balance in allowing themselves room to carry out useful research whilst at the same time making a real effort to inform individuals.

Another option in terms of a legal base is to make use of the scientific research exception. This wide-ranging exception is available to individuals or organisations conducting scientific research. Its most obvious advantage is that using it means that researchers are not compelled to obtain consent from data subjects. Use of the research exception also provides more freedom in terms of complying with the purpose of processing at the time the data was collected (important in the use of secondary data). Further advantages include the fact that a number of data subject rights may not apply when processing is based on this legal ground. This may include the right to erasure and the right to demand that one's data is no longer processed. Using this legal ground cannot however be

of health research: getting the balance right. *Quaterly Journal of Medicine,* 107, 3-5.p4.

[124] Jensen, P., Jensen, L. & Brunak, S. 2012. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics,* 13, 395-405 p395, 397, 399.

viewed as a *carte blanche* option as it comes with important conditionality. Most importantly, it can only be used when it is necessary i.e. where consent would not be suitable because it would prevent the scientific research in question. Researchers can only therefore opt to use this legal ground where there is a genuine reason making consent unsuitable. Even where this is the case however, the GDPR requires that adequate measures be taken to protect the security and privacy of data subjects. Member States may add further requirements in their law to the same end. The result is that even where such an option is both available and suitable, it may entail undertaking a number of measures that may have a negative impact on researchers. This may be either in terms of cost (in both a temporal and pecuniary sense) where complex administrative measures have to be engaged in, or in terms of the ambitiousness of the research goals where data must be pseudonymsed to a high degree.

Whilst the authors of this paper accept that consent should remain the default option, (i.e. to be used where suitable), they are of the opinion that it is important to recognise that it may not always be suitable. This may be the case for example where the cognitive or educational background of data subjects mean that they are unlikely to be able to understand the processing applied to their data. Other examples include situations where it is difficult to track all potential data subjects (e.g. where data is to be used from a previous experiment). In some instances ethics bodies seem unwilling to utilise the scientific research option and seem to prefer consent, even where this may be unfeasible given the research that is being proposed. Given that this is often not feasible, this has in a number of instances pushed researchers into a consent/anonymise dichotomy where the option of processing data under the scientific research ground is not given due consideration. Unfortunately, this may in reality mean that research is hampered (i.e. where data is unnecessarily anonymised) or that either consent or anonymisation (which is arguably impossible with large amounts of sensitive data) is used inappropriately. The authors of this paper would, call both for better research into this phenomenon and also for an improved appreciation of the suitability of the scientific research option, especially given the conditionality that accompanies it (and which can arguably provide adequate privacy protection in many instances). In doing so important and innovative research can be facilitated, whilst at the same time protecting the rights of research participants and data subjects in an appropriate way.

REFERENCES

Al Aziz MM, Hasan MZ, Mohammed N, Alhadidi D. Secure and efficient multiparty computation on genomic data. Proceedings of the 20th International Database Engineering & Applications Symposium 2016:278–83.
Aldhouse F. Anonymisation of personal data - A missed opportunity for the European Commission. Comput Law Secur Rev 2014;30:403–18.
Amin V, Behrman JR, Spector TD. Does more schooling improve health outcomes and health related behaviors? Evidence from U.K. twins. Econ Educ Rev 2013;35. doi:10.1016/j.econedurev.2013.04.004.
Andrews L. Social, legal, and ethical implications of genetic testing. Wshington (DC): National Academic Press; 1994.
Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. The 1000 Genomes Project, C. A global reference for human genetic variation. Nature 2015;526:68–74.
Boddington P, Curren L, Kaye J, Kanellopoulou N, Melham K, Gowansa H, et al. Consent forms in genomics: the difference between law and practice. Eur J Health Law 2011;18:491–519.
Bohannon J. Genealogy databases enable naming of anonymous DNA donors. Science 2013;339. doi:10.1126/science.339.6117.262.
Butler J. The future of forensic DNA analysis. Phil Trans R Soc 2015. doi:10.1098/rstb.2014.0252.
Cai R, Hao Z, Winslett M, Xiao X, Yang Y, Zhang Z, et al. Deterministic identification of specific individuals from GWAS results. Bioinformatics 2015;31:1701–7.
Carter P, Laurie G, Dixon Woods M. The social licence for research: why care data ran into trouble. J Med Ethics 2015. doi:10.1136/medethics-2014-102374.
Chassang G. The impact of the EU general data protection regulation on scientific research. Ecancermedicalscience 2017;11:709.
Clayton E, Mizielinska S, Edgar J, Nielsen T, Marshall S, Norona F, et al. Frontotemporal dementia caused by CHMP2B mutation is characterised by neuronal lysosomal storage pathology. Acta Neuropathol (Berl) 2015;130:511–23.
Consortium TUK. The UK10K project identifies rare variants in health and disease. Nature 2015;526:82–90.
Dawn Teare M, Barrett JH. Genetic linkage studies. Lancet North Am Ed 2005;366:1036–44.
De Hert P, Gutwirth S. Privacy and the criminal law. In: Claes E, Duff A, Gutwirth S, editors. Privacy, data protection and law enforcement. opacety of the individual and transparency of the power. Antwerp: Oxford: Intersentia; 2006.
Deribe K, Beng A, Cano J, Njouendo A, Fru-Cho J, Awah A, et al. Mapping the geographical distribution of podoconiosis in Cameroon using parasitological, serological, and clinical evidence to exclude other causes of lymphedema. PLoS NeglTrop Dis 2018 https://doi.org/10.1371/journal.pntd.0006126.
Dubois L, Kyvik K, Girard M, Tatone-Tokuda F, Pérusse D, Hjelmborg J, et al. Genetic and environmental contributions to weight, height, and bmi from birth to 19 years of age: an international study of over 12,000 twin pairs. PLoS One 2012 https://doi.org/10.1371/journal.pone.0030153.
Fears R, Brand H, Frackowiak R, Pastoret P, Souhami R, Thompson B. Data protection regulaiton and the promotion of health research: getting the balance right. Q J Med 2014;107:3–5.
Friedewald M, Hallinan D. Open consent, biobanking and data protection law: can open consent be 'informed' under the forthcoming data protection regulation? Life Sci Soc Policy 2015;11:1–36.
Ghani N, Hamid S, Udzir I. Big data and data protection - issues with purpose limitation principle. Int J Adv Soft Comput Appl 2016;8:116–21.
Gutwirth S, Leenes R, De Hert P, Poulletn Y. European data protection: in good health?. Springer; 2012.
Gymrek M, Mcguire A, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science 2013;339:321–4.
Hallinan D, Friedewald M, De Hert P. Genetic data and the data protection regulation: anonymity, multiple subjects, sensitivity and a prohibitionary logic regarding genetic data? Comput Law Security Rev 2013;29:317–29.
He KY, Ge D, He MM. Big data analytics for genomic medicine. Int J Mol Sci 2017;18:412.

Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. Genom Inform 2012;10:117–22.
Hudjashov G, Karafet T, Lawson D, Downey S, Savina O, Sudoyo H, et al. Complex patterns of admixture across the indonesian archipelago. Mol Biol Evol 2017;34:2439–52.
Jensen P, Jensen L, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13:395–405.
Kaye J. The tension between data sharing and the protection of privacy in genomics research. Annu Rev Genom Hum Genet 2012;13:415–31.
Khaled E, Alvarez C. A critical appraisal of the article 29 working party opinion 05/2014 on data anonymization techniques. Int Data Privacy Law 2015;5:73–87.
Kohane I. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet 2011;12:417–28.
Konigsberg L, Frankenberg S. Bayes in biological anthropology. Am J Phys Anthropol 2013;57:153–84.
Livy A, Sayhean L, Jagdish C, Hanis N, Sharmila V, Wee Ler L, et al. Evaluation of quality of DNA extracted from buccal swabs for microarray based genotyping. Indian J Clin Biochem 2012;27:28–33.
Lugg W, Griffiths J, Van Rooyen A, Weeks A, Tinglet R. Optimal survey designs for environmental DNA sampling. Methods Ecol Evolut 2017. doi:10.1111/2041-210X.12951.
Lynskey O. Deconstructing data protection: the 'Added-Value' of a right to data protection in the EU legal order. Int Comp Law Q 2014;63:569–97.
Lyu, H., Huan, J., Zhimin, H. & Liu, B. 2018. Epigenetic mechanism of survivin dysregulation in human cancer.
Malaspinas AS, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic history of Aboriginal Australia. Nature 2016;538:207–14.
Malgieri G, Comandé G. Sensitive-by-distance: quasi-health data in the algorithmic era. Inf Commun Technol Law 2017;26:229–49.
Mantovani E, Quinn P. mHealth and data protection – the letter and the spirit of consent legal requirements. Int Rev Law, Comput Technol 2013 http://dx.doi.org/10.1080/13600869.2013.801581.
Mcguire AL, Beskow LM. Informed consent in genomics and gentic research. Annu Rev Genom Hum Genet 2010;11:361–81.
Mcguire AL, Hamilton JA, Lunstroth R, Mccullough LB, Goldman A. DNA data sharing: research participants' perspectives. Genet Med 2008;10:46–53.
Mead S, Uphill J, Beck J, Poulter M, Campbell T, Lowe J, et al. Genome-wide association study in multiple human prion diseases suggests genetic risk factors additional to PRNP. Hum Mol Genet 2012;21:1897–906.
Moltke I, Albrechtsen A, Hansen T, Nielsen F, Nielsen R. A method for detecting IBD regions simultaneously in multiple individuals–with applications to disease genetics. Genome Res 2011;21:1168–80.
Niemiec E, Howard H. Ethical issues in consumer genome sequencing: use of consumers' samples and data. Appl Transl Genet 2016;8:23–30.
Nuzzo A, Riva A, Bellazi R. Phenotypic and genotypic data integration and exploration through a web-service architecture. BMC Bioinform 2009 https://doi.org/10.1186/1471-2105-10-S12-S5.
Palazzo A, Gregory T. The case for junk DNA. PLos Genet 2014 https://doi.org/10.1371/journal.pgen.1004351.
Quinn P. The anonymisation of research data — a pyric victory for privacy that should not be pushed too hard by the EU data protection framework? Eur J Health Law 2017;24. doi:10.1163/15718093-12341416.
Roche M, Berg J. Incidental findings with genomic testing: implications for genetic counseling practice. Curr Genet Med Rep 2015;3:166–76.
Roewer L. DNA fingerprinting in forensics: past, present, future. Investig Gent 2013 https://doi.org/10.1186/2041-2223-4-22.
Schmidt H, Callier S. How anonymous is 'anonymous' ? Some suggestions towards a coherent universal coding system for genetic samples. J Med Ethics 2012;38. doi:10.1136/medethics-2011-100181.
Shabani M, Borry P. Rules for processing genetic data for research purposes in view of the new EU general data protection regulation. Eur J Hum Genet 2017.
Shoenbill K, Fost N, Tachinardi U, Mendonca E. Genetic data and electronic health records: a discussion of ethical, logistical and technological considerations. J Am Med Inform Assoc 2014. doi:10.1136/amiajnl-2013-001694.
Takahashi J, Pinto L, Vitaterna M. Forward and reverse genetic approaches to behavior in the mouse. Science 1994;264:1724–33.
Tischkoff S, Verreli B. Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genom Hum Genet 2003;4:293–340.
Zheng-Bradley X, Flicek P. Applications of the 1000 genomes project resources. Briefings Funct Genom 2017;16:163–70.
Zhong M, Zhang Y, Lange K, Fan R. A cross-population extended haplotype-based homozygosity score test to detect positive selection in genome-wide scans. Stat Interface 2011;4:51–63.