Protection of genomic data and the Australian Privacy Act: when are genomic data 'personal information'?

Minna Paltiel*, Mark Taylor ** and Ainsley Newson ***,****

Key Points

- 'Personal information', protected under the Australian *Privacy Act 1988* (Cth), is 'about an identified individual or an individual who is reasonably identifiable' (S.6), so the legal assessment of 'identifiability' shapes the protection of genomic data under the *Privacy Act*.
- Not all genomic data are captured by the statutory definitions of 'genetic information' in the *Privacy Act*; however, genomic data that do not fit the definition may still be protected if they are about an identifiable individual.
- In applying the legal test of identifiability to genomic data, the interaction between the data and the data environment must be examined. Overemphasis on particular features of genomic data, such as 'rareness' or 'uniqueness', may lead to a misapplication of the *Privacy Act*.
- Whether genomic data are personal information is primarily a matter of the opportunities and likelihood of linking the genomic data in question with other data available in the data environment.

Introduction

Sharing genomic data promises great benefits for health research as well as clinical diagnoses and management. However, appropriate sharing is reliant upon privacy concerns being effectively addressed.[1] In Australia, the first step toward understanding the protection of genomic data under federal law (specifically the Australian *Privacy Act 1988* (Cth) (*Privacy Act*)) is understanding when genomic data are protected by that Act. The *Privacy Act* covers, inter alia, commonwealth government agencies and private sector health service providers.[2] Collection, use, and disclosure of data by such entities are, however, only regulated by the Act in so far as the data themselves also fall within the material scope of the legislation. This article considers the relationship between the term 'genomic data', as it might be used in scientific (or lay) conversation, and the concept of 'genetic information' provided by law. This is important for at least three reasons: (i) those subject to the *Privacy Act* need to be able to confidently navigate their responsibilities, such as knowing when consent to sharing is required; (ii) understanding current controls is a prerequisite for meaningful external critique (and this is particularly important at present, given that the *Privacy Act* is under review); and (iii) while legislation that applies to state public sector agencies is generally[3] distinct from the *Privacy Act*,[4] there are similarities that

* Minna Paltiel, Melbourne Law School, The University of Melbourne, Melbourne, VIC, Australia
** Mark Taylor, Melbourne Law School, The University of Melbourne, Melbourne, VIC, Australia
*** Faculty of Medicine and Health, Sydney School of Public Health, The University of Sydney, Sydney, NSW, Australia
**** Australian Genomics, Melbourne, VIC, Australia

1 Appropriate sharing is a particular consideration when there is no clear consent attached to the data in question.

2 Note that generally what are defined as 'small businesses' under the *Privacy Act* are not considered APP entities. 'APP entities' are agencies or organizations which are bound by the Australian Privacy Principles, found in the *Privacy Act* sch 1. Small businesses are excepted. They are not APP entities under the Act and are not bound by it. However, this exception does not apply to health service providers or entities holding health information (other than with regard to employee records) (*Privacy Act*, s 6D(4)).

3 In the case that both Commonwealth and State privacy legislation apply to an entity, and an inconsistency exists, s 109 of *Commonwealth of Australia Constitution Act* requires that where there is inconsistency between Commonwealth and State law, Commonwealth law prevails to the extent of the inconsistency.

4 See in Victoria the *Privacy and Data Protection Act 2014* (Vic) and the *Health Records Act 2001* (Vic); in Queensland, the *Information Privacy Act 2009* (Qld); in New South Wales, the *Privacy and Personal information Protection Act 1998* (NSW) and the *Health Records and Information Privacy Act 2002* (NSW); and, in Tasmania, the *Personal Information Protection Act 2004* (Tas).

extend the relevance of the question: When are genomic data 'personal information' under the *Privacy Act*?

Common to almost all privacy legislation in Australia is the fact that the material scope of statutory privacy protection extends only to the 'handling of personal information'.[5] The *Privacy Act* defines a number of subcategories of personal information such as 'sensitive information', 'health information', and 'genetic information' (see the section, '*Privacy Act* definitions of protected 'personal information' and genomic data' below). While understandable, we argue that it should not be assumed that genomic data will be captured by these definitions. Indeed, a key claim of this article is that not all genomic data are 'personal information' or 'genetic information' as legally defined.

In this article, we explain why the first and essential question, when assessing whether genomic data are subject to privacy restrictions, is whether the genomic data are 'personal information'. In relation to this, the current *Privacy Act* definition of personal information requires a two-pronged test: data must be 'about' an individual, and that individual must be 'identified' or 'reasonably identifiable'.[6] In this article, we focus on the second prong of this test and consider when genomic data fulfil the 'identifiability' requirement. We adopt this focus not only because other work, and our own experience, has indicated this aspect of the test to be the subject of confusion in practice, but also because there is indication that the current requirement for data to be 'about' a person may be reformed as a result of the current *Privacy Act* review.[7] Even if the definition of personal information is amended, there is good reason to think that identifiability will remain a key consideration.[8]

Before engaging in a detailed analysis of the conceptual overlap between genomic data and 'personal information', we will acknowledge as understandable any pre-theoretical assumption that all genomic data are about an 'identified' or 'identifiable' individual. After all, genomic data have been described as, by their nature, 'strongly identifying'[9] and scholars have commented on the 'uniqueness' of every individual's genomic data.[10] However, we suggest that from a privacy law perspective the identificatory potential of genomic data, in terms of scientific possibility (or lay understanding), does not answer the question of whether data are *legally* 'identifiable' in all the circumstances.[11] The legal test and threshold of identifiability is constructed according to the rules by which legal knowledge is constructed; it is not a scientific question.

To explain the relationship between genomic data and legal concepts of 'personal information' and 'genetic information', it is useful to start by establishing some distinctive scientific characteristics of genomic data. We do this in the first section, 'Categories of genomic data'. This is followed in the second section, '*Privacy Act* definitions of protected 'personal information' and genomic data' with an analysis of *Privacy Act* definitions, thus allowing us to conclude that section with a consideration of the (non) alignment between scientific and legal definitions. In the third section, 'The *Privacy Act* test of reasonable identifiability', we set out the assessment of legal identifiability in the *Privacy Act*, according to the guidance provided by the Office of the Australian Information Commissioner (OAIC), relevant case law, and academic literature. In this section, we discuss the 'evaluative' nature of the required assessment and the significance of the relationship between features of genomic data and the data processing environment. In the fourth section, 'Identifiability of genomic data', we analyse these elements of genomic data and the data environment relevant to the evaluation of identifiability. In the concluding section, we summarize the key point of the analysis in the previous sections, and the processes required in assessing the identifiability of particular genomic data and their legal status with regard to *Privacy Act* protections.

A preliminary note should be made of the (often interchangeable) use of the terms 'genetic' and 'genomic'. Put simply, 'genetics' refers to 'the study of single genes and their effect within an organism', while 'genomics' refers to the 'study of many genes simultaneously'.[12] In Australia, the National Statement on Ethical Conduct in

5 *Privacy Act*, s 2A.

6 *Privacy Act*, s 6.

7 Australian Attorney-General's Department, *Privacy Act Review Discussion Paper* (October 2021) 21–24.

8 The question of when genomic data is 'about' an individual, and the consequence of extending that to data which 'relates to an individual,' are worthy of separate consideration.

9 See eg, Thomas Finnegan and Alison Hall, *Identification and Genomic Data* (PHG Foundation 2017) 12. See also Lawrence O Gostin, 'Genetic Privacy' (1995) 23(4) Journal of Law, Medicine and Ethics 320, 324.

10 See eg, Finnegan and Hall ibid 12; Gostin ibid 324; Muhammad Naveed and others, 'Privacy in the Genomic Era' (2015) 48(1) ACM Computing Surveys 1, 2, 11. Writers also note the 'mystique' regarding public perception of genomic data. (See Finnegan and Hall (n 9) 12; Naveed and others, 'Privacy in the Genomic Era' (2015) 48(1) ACM Computing Surveys 1, 2, 11.) Even those arguing against genetic exceptionalism noted the public perception that genetic information is different—more powerful, 'mysterious', and sensitive—to other forms of health information. See Thomas H Murray, 'Genetic Exceptionalism and "Future Diaries": Is Genetic Information Different from Other Medical Information' in Mark Rothstein (ed), *Genetic Secrets* (Yale University Press 1997); Lawrence O Gostin and Games G Jr Hodge, 'Genetic Privacy and the Law: An End to Genetics Exceptionalism' (1999) 40 Jurimetrics 21, 36.

11 See Finnegan and Hall (n 9) 11.

12 Erin Turbitt and Barbara B Biesecker, 'A Primer in Genomics for Social and Behavioral Investigators' (2020) 10(2) Translational Behavioral Medicine 451, 452.

Human Research ('National Statement'),[13] in addressing ethical considerations in the design and conduct of genomic research, defines genomic research as including the 'full scope of genetic research'.[14] Privacy legislation, guidelines, and codes referred to in this article use the term 'genetic information'. However, the definitions of 'genetic information' (which are set out in the third section, '*Privacy Act* definitions of protected 'personal information' and genomic data') broadly encompass genomic data as the term is used in either a lay or scientific sense. In this article, the term 'genetic information' is used when discussing the legislative definition of genetic information. In contrast, we associate the term 'genomic data' with scientific ontologies. This allows us to consider the extent to which the boundaries between 'genetic information' and 'genomic data' are coterminous when considered from legal or scientific perspectives.

Categories of genomic data

'Genomic data' may describe data with a wide range of characteristics[15] and is referred to with varying degrees of specificity or generality. It is sometimes used to refer narrowly to data produced by genetic tests such as predictive or carrier tests, at other times the term extends to incorporate information such as family history and pedigree data as part of 'genomic content', or even more broadly to include 'supporting' clinical and administrative data.[16] For clarity, in this analysis, we use the term 'genomic data' to refer only to DNA sequence data derived from molecular testing (such as data obtained using next-generation sequencing). This includes data derived from whole-genome sequencing (WGS) and whole-exome sequencing (WES) as well as variant data—single-nucleotide variants (SNV), rare variants, and variants of uncertain significance.[17] Our review thus excludes certain genomic data, such as data derived from family history or other kinds of testing, such as cytogenic testing. Other types of data attached to the genomic data to support its interpretation are referred to as 'supporting data'. The role and significance of 'supporting data' in the assessment of identifiability are discussed in the following sections.

Types of genomic data

Raw and interpreted data

'Sequence read data', which includes WGS and WES data as well as data relating to SNPs, are generated through sequencing technology and may be considered 'raw' genomic data. It includes the biological sequence data and the 'quality score', which indicates the probability that a 'read' (denoting a DNA base pair) was 'called' correctly.[18]

Interpreted data refers to data that has been analysed or annotated. 'Analysed genomic data' are produced by aligning sequences to a reference genome and making 'variant calls' which identify differences between the sample sequence and the reference genome.[19] Identified differences between the aligned reads are generally written to a variant call format (VCF).[20] VCF files may include 'variant identifiers' which are unique combinations of letters and numbers assigned to genes, variants, or proteins, used to identify them.[21]

'Annotated data' refers to information about identified variants, relevant to variant classification. This may involve using the VCF file together with known variant databases (such a ClinVar[22] and Shariant)[23] to produce an annotated VCF, which would include details of external data, such as phenotype obtained from phenotype–genotype mapping.[24]

13 Australian Research Council and Universities Australia, *National Statement on Ethical Conduct in Human Research 2007* (National Health and Medical Research Council first published 2007, updated 2018).

14 Ibid ch 3.3, 101.

15 See discussion in Finnegan and Hall (n 9) 4.

16 See Queensland Government, *Blueprint for a National Approach to Genomic Data Management* (Queensland Health, Australia 2020) 39, 41.

17 Turbitt and Biesecker (n 12) 453–54.

18 See Queensland Government (n 16) 41-42. Sequencing quality scores indicate the probability that the base (one of the four phosphate bases of the DNA - adenine, thymine, cytosine and guanine, whose initials, ATCG, 'spell' the genetic code) was called correctly. The quality score is assigned by a 'phred-like algorithm' similar to that of the original Sanger sequencing. (See Illumina, 'Measuring Sequencing Accuracy' <https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html> accessed 6 August 2022; 23andMe, 'Genetics 101 (Part 1 of 5): What are Genes?' (18 April 2012) <https://www.youtube.com/watch?v=ubq4eu_TDFc> accessed 6 August 2022.)

19 Queensland Government (n 16) app 94; European Molecular Biology Laboratory – European Bioinformatics Institute, 'Variant Identification and Analysis' <https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction/variant-identification-and-analysis/> accessed 23 February 2021. Reported variants are then interpreted by the referring clinician (variant interpretation).

20 Queensland Government (n 16) app 94; European Molecular Biology Laboratory – European Bioinformatics Institute (n 19).

21 European Molecular Biology Laboratory – European Bioinformatics Institute (n 19).

22 National Library of Medicine, 'ClinVar' <https://www.ncbi.nlm.nih.gov/clinvar/> accessed 6 August 2022.

23 Australian Genomics, 'Shariant' <https://www.australiangenomics.org.au/tools-and-resources/shariant/> accessed 6 August 2022.

24 Queensland Government (n 16) apps 94–95; European Molecular Biology Laboratory – European Bioinformatics Institute (n 19).

Sequence read and analysed data: volume and richness

WGS is a read of a person's entire nuclear genome. WES is a read of only 'exons', those regions of the DNA which code for proteins. WGS and WES data may be stored in raw form to be reanalysed and reinterpreted for purposes other than the original purpose of testing. We currently understand each individual to have unique whole sequence read data.[25]

A difference in the DNA sequence affecting a single base pair in a genome is an SNV. Common SNVs (occurring in more than 1 per cent of the population) are called single nucleotide polymorphisms (SNPs).[26] SNPs are the most common type of variant, and most do not lead to any observable differences between people.[27] SNPs are studied in combination to create a profile, which can be used to identify combinations of SNPs that indicate a higher risk of developing certain diseases.[28]

The terms 'rare variant' or 'novel variant' are used to describe variants found in less than 1 per cent of the population.[29] The term '*de novo* mutation' has also been used to describe a variant that arises in one individual, but is not, for example, present in the somatic genome of the parents.[30] *De novo* variants are part of what makes that person's physiology unique.[31] On the other hand, some variants will be shared between family members, or found at higher levels of frequency within particular populations.[32] The relative scarcity or abundance of annotated data will depend on the frequency of both the variant and the specific annotation, with the latter itself dependent on the availability and use of particular variant databases.

Genomic data are also sometimes distinguished from other data types with regard to the size of the data set and its subsequent volume and richness.[33] It is important to recognize that the volume and richness of a data set is distinguishable from 'volume and richness' of specific genomic data and that both may be hugely variable.

'Supporting data'

As mentioned above, the scope of 'genomic data' is sometimes taken to include a range of 'supporting data' other than that generated by genomic testing. Genomic data are generally collected, stored, and shared with other types of data such as 'curation data', and data from external sources such as 'reference genome', publications, details of prevalence in the relevant population, known biological functions, and genotype/phenotype correlations, etc. 'Supporting data' may also include clinical observations and metadata pertaining to the collection of the data. Supporting data may facilitate linkages between data sets and associations with individuals. As we explain in the following discussion, linkages which are drawn between genomic data and various supporting data, and potentially between supporting data and other information in a particular data environment, are likely to play an important role in the legal assessment of 'identifiability'.

Privacy Act definitions of protected 'personal information' and genomic data

Privacy Act definition of 'personal information' and identifiability

In the *Privacy Act*, *personal information* is stated in section 6(1) to mean:

> information or an opinion about an identified individual, or an individual who is reasonably identifiable:
>
> (a) whether the information or opinion is true or not; and
>
> (b) whether the information or opinion is recorded in a material form or not.

This definition of 'personal information' has applied since the Act was amended in 2012.[34] The repealed

25 An exception to this would be identical twins, although identical twins may occasionally have different reads where a mutation occurs in an embryo after the twin embryos have split.

26 Harvard University, 'An Introduction the Human Genome' (20 May 2017) <https://www.youtube.com/watch?v=jEJp7B6u_dY> accessed 23 February 2021.

27 23andMe, 'Genetics 101 (Part 2 of 5): What are SNPs?' (18 April 2012) <https://www.youtube.com/watch?v=tJjXpiWKMyA> accessed 6 August 2022.

28 Turbitt and Biesecker (n 12) 453.

29 See Aude Saint Pierre and Emannuelle Genin, 'How Important are Rare Variants in Common Disease' (2014) 13(5) Briefings in Functional Genomics 353. The threshold of 1 per cent of the population has been called 'arbitrary' by Dudley and Karczewski who wrote that this may be sufficiently rare to detect variants contributing to the risk of an adverse drug effect, but is not sufficiently rare to determine the cause of an 'extremely rare disease' (eg, Miller Syndrome). See Joel T Dudley and Konrad J Karczewski, *Exploring Personal Genomics* (OUP 2013) 199.

30 Dudley and Karczewski ibid.

31 Dudley and Karczewski (n 29) 199, 200–01. Note that every individual carries around 30–100 *de novo* variants.

32 Dudley and Karczewski (n 29) ch 10, 199–220.

33 'Richness' of data refers to the capacity of the data to reveal details and complexities of the matter being studied. Alan RA Aitken and others, 'A Role for Data Richness Mapping in Exploration Decision Making' (2018) 99 Ore Geology Reviews <https://doi.org/10.1016/j.oregeorev.2018.07.002> accessed 6 August 2022.

34 The definition of 'personal information' was amended by the *Privacy Amendment (Enhancing Privacy Protection) Act 2012* (Cth), sch 1, item 36. The new definition came into effect in March 2014. The amendment was based upon the recommendation of the Australian Law Reform

definition referred to 'an individual whose identity is apparent or can be reasonably ascertained *from* the information or opinion'.[35] The amendment is significant not only as it extended the concept of personal information, but also because it is an amendment that has not (yet) been adopted by state legislation. Definitions of 'personal information' in privacy and health record legislation in Victoria, NSW, Queensland, and Tasmania maintain the requirement that the identity of a person must be apparent or reasonably ascertained 'from the information or opinion'.[36] The difference could be significant in some scenarios, particularly in the case of data sharing. For example, if genomic data were shared open access by a state government agency on the basis that it was not personal information, then while the agency may be compliant with its governing legislation, those accessing and processing the data, if bound by the *Privacy Act*, may be accessing personal information according to that Act's definition.

The requirement that personal information is about an identified or identifiable individual is central to both the previous and the amended *Privacy Act* definitions of 'personal information'. It is, however, important to underline that under the current definition the relevant question is whether the genomic data are about a person who is identified or identifiable. The question is answered with reference to, but goes beyond, the question of whether the genomic data could be used to identify them. We return to this point later when we apply the legal test of identifiability to genomic data.

It is also important to recognize that the question of whether data are genetic information, for the purposes of an application of the *Privacy Act*, should be answered via a different methodology to that used in a scientific ontology. The category and concept of genomic data will be constructed, and contested, in scientific (and lay) discourse using different rules of knowledge formation than are relied upon in legal argument. This can easily cause confusion and is a key point to convey to data custodians.

'Genetic information' in the Privacy Act

The *Privacy Act* considers 'genetic information' to be 'sensitive information'. 'Sensitive information' is a subset of personal information and is considered to pose particularly adverse consequences for the data subject or another if mishandled.[37] As such, it is afforded higher levels of protection. 'Health information' is a subcategory of 'sensitive information' and includes 'genetic information about an individual in a form that is, or could be, predictive of the health of the individual or a genetic relative of the individual'.[38] 'Genetic information' that is not health information is also, separately, included as a subcategory of 'sensitive information'.[39] The *Privacy Act* and explanatory material do not explain what is meant by 'genetic information that is not health information'; however, this category of genetic information would presumably include, for example, genomic data used forensically to identify an individual, which *per se* would relate to an 'identified or reasonably identifiable individual'.[40]

In addition to the distinction of being included twice within the sub-category of sensitive information (as health information and as genetic information that is not health information) there is another curiosity in the Act's definition of 'genetic information'. Other types of information defined as 'sensitive information', such as political opinions, religious belief, and sexual orientation, are expressly required under the Act to be *also* 'personal information'.[41] This is not so with 'health information' or 'genetic information that is not otherwise health information'. The definition of 'health information'[42] does require that other types of health information be 'personal information', but this is not expressly

Commission, 'For Your Information: Australian Privacy Law and Practice' (ALRC Report 108, 12 August 2008), para 6.55.

35 Emphasis added.

36 *Privacy and Data Protection Act 2014* (Vic) s 3. Similar definitions appear in the *Health Records Act 2001* (Vic) s 3; *Information Privacy Act 2009* (Qld) s 12; *Privacy and Personal information Protection Act 1998* (NSW) s 4(1); *Health Records and Information Privacy Act 2002* (NSW) s 5; *Personal Information Protection Act 2004* (Tas) s 3.

37 OAIC, *Australian Privacy Principles Guidelines, Privacy Act 1988* (first published February 2014, combined July 2019) para B.141; OAIC, *What is Personal Information*? (May 2017) 4. See also OAIC, 'What is Personal Information?' (5 May 2017) <https://www.oaic.gov.au/privacy/guidance-and-advice/what-is-personal-information> accessed 6 August 2022.

38 *Privacy Act* s 6FA(d).

39 *Privacy Act* s 6.

40 Similar references to genetic data are made in the *EU General Data Protection Regulation (GDPR):* Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Dir 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1 which includes 'genetic data' as an element included in 'data concerning health' (recital 35), and also as distinct category of 'personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question' (art 4(13), recital 34). This second category would capture genetic data processed expressly for the purpose of identifying a person, (for example to establish parentage) (art 9(1), recital 34, 35).

41 *Privacy Act* s 6.

42 *Privacy Act* s 6FA.

stated for 'genetic information'.[43] Interpreting these provisions in context, and taking into account the stated objects of the Act, we do not think anything can turn on this quirk of drafting. Genetic information (whether health information or not) will only fall within scope of the *Privacy Act* if it is also personal information.[44] To understand these provisions otherwise would not only run counter to principles of statutory interpretation,[45] but would leave the statutory concept of 'genetic information' almost entirely unbounded: anything that might be described as genetic information could fall subject to the provisions of the Privacy Act. This could include any genetic facts or statistics in any context, including those relating to all living persons. We suggest it is unhelpful, and most unlikely to be legally sound, if drafting is understood to imply a categoric difference between genetic information and other kinds of sensitive data with regard to the requirement that it is personal information, and the question of identifiability. The current *Privacy Act* review may present an opportune moment to address this anomaly and to put beyond doubt that the only genetic information within scope of the *Privacy Act* is personal information.

Taken together, this discussion shows that what we have called 'genomic data' may fall under at least three categories, only two of which are recognized as 'genetic information' in law in Australia: genetic information that is health information, genetic information that is not health information but remains sensitive information, and genomic data that are not personal information at all (and as such would not be subject to the *Privacy Act*).

Before moving on to consider when genomic data will be considered personal information under the *Privacy Act* we draw attention to one further notable feature of the Act's definition of genetic information. Genetic information is only a subset of health information when it is '. . .in a form that is, or could be, *predictive* of the health' of the person providing the information, or a genetic relative.[46] The element of predictiveness is ambiguous and may not align with the terms and meaning used in the field of genetics. In the study of genetics, 'predictive' testing and 'predictive' data refers specifically and narrowly to testing an asymptomatic individual in order to predict a future risk of disease.[47] It seems unlikely that the statutory definition was intended only to apply to genetic information about asymptomatic individuals. We do not further consider this point here, as any genetic information that is not health information is still covered by the Act, but it is potentially significant and helps illustrate our broader point relating to legal definitions not aligning with scientific terms and meanings.

The *Privacy Act* test of reasonable identifiability

What does it mean for information to be 'identifiable'?

According to OAIC guidance, information is about an identified person when a particular individual[48] within a group of persons can be 'distinguished' from the others via a link between the information in question and a particular person.[49] When data includes identifiers that directly indicate an 'identified' individual, the information will, on the face of it, be about an individual distinguished from others and will be personal information. The issue of identifiability is less clear when personal information is about an individual who is 'reasonably identifiable'. 'Reasonable identifiability' is a complex criterion and there are no strict rules provided in statute or soft law for its determination.[50] However, we can use case law and OAIC guidance to discern two important aspects of the assessment of identifiability: (i) that reasonable identifiability is context dependent, and

43 *Privacy Act* s 6FA(a) and (d) respectively.

44 *Privacy Act* s 2a. OAIC guidance also makes it clear that only data that is 'personal information' is subject to *Privacy Act* regulation, see OAIC, *Australian Privacy Principles Guidelines, Privacy Act 1988* (n 37) [B.60]; See also Department of Health, *Framework to Guide the Secondary Use of My Health Record System Data* (May 2018) ('*Secondary Use Framework*') 19–21.

45 We apply here the maxim of *noscitur a socis*, according to which words in statute are to be interpreted according to (or 'coloured by') their surrounding provisions (*R v Ann Harris* (1836) 7 Car &P 446; 173 ER 198). We also refer to the principles of statutory interpretation provided in the *Acts Interpretation Act 1901* (Cth), including the requirement to interpret statute according to its purpose or object. Sections 2A(c), (d) of the *Privacy Act*, refer to the regulation and promotion of responsible and transparent handling of 'personal information'.

46 *Privacy Act* s 6FA(d) (emphasis added).

47 See James P Evans, Cécile Skrzynia and Wylie Burke, 'The Complexities of Predictive Genetic Testing' (2001) 322(7293) British Medical Journal 1052 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1120190/#:~:text=Predictive%20genetic%20testing%20is%20the,predict%20future%20risk%20of%20disease> accessed 6 August 2022.

48 The *Privacy Act*, s 6 defines 'individual' to mean 'a natural person'.

49 OAIC, *What is Personal Information*? (n 37) 8.

50 The OAIC directs that 'there is no exact formula for assessing when information will be reasonably identifiable, and it can sometimes be difficult to draw a bright line between de-identified information and personal information' (OAIC, *De-identification and the Privacy Act* (March 2018) 8).

(ii) that it turns upon the 'reasonableness' of the prospect of identification.

Courts and tribunals have recognized the importance of context in assessing identifiability, finding that documents and records containing no obvious identifying features may nevertheless 'take on the quality' of identifiability through the context in which they are held.[51] Rangiah J in *Baptist Union of Queensland – Carinity v Roberts and ors*[52] emphasized that 'reasonably identifiable' information includes information from which the identity of the individual may be reasonably identifiable using information held by any other entity.[53] This accords with what we have said above regarding the amended *Privacy Act* definition of 'personal information' and the fact that a person need not be identified *from* the data: it is enough for data to be 'about' an individual who can be distinguished from others in the context due to a link between the information and a particular person.[54]

Therefore, the standard in the test of identifiability is whether the genomic data, on its own or linked with other information, are about a distinguishable individual, whether or not the genomic data are used, or will be used, to identify that individual.

In *Baptist Union*, Rangiah J also affirmed that information is 'reasonably identifiable' only where it is reasonably practicable to identify the individual, taking into account the 'likelihood, cost, difficulty and practicability' of identification.[55] 'Reasonable identifiability' is thus not a theoretical question, but a practical test. The technical possibility of identifying an individual is not sufficient to qualify data as personal information. In situations where identification is unlikely or impracticable to achieve, the information will not be personal information.

An evaluative process

The evaluative nature of determinations of identifiability was described in *Privacy Commissioner v Telstra Corporation Ltd.*[56] Although the Court did not address the meaning of 'reasonable identifiability' directly, Kenny and Edelman JJ wrote that 'a determination of whether the identity can reasonably be ascertained will require an evaluative conclusion'.[57] The assessment of 'reasonable identifiability' thus requires consideration of many interrelated factors relating to both the genomic data and to the data processing context, and is undertaken on a case-by-case basis, taking into account all the circumstances. An illustration of such an approach is given by the OAIC, which suggests that in assessing 'reasonable identifiability' one should consider: the nature and quantity of information; the circumstances of its receipt; the people or organization who hold or can access the information; other information available to the entity holding the information; the potential ability of the entity holding the information, using resources available to it, to identify the person to whom it relates and where this is possible 'the practicability, including the time and cost involved' in doing so; and, if the information is shared publicly, whether a 'reasonable member of the public' would be able to use it to identify the individual.[58]

These considerations bear upon the 'reasonableness', that is, the 'likelihood, cost, difficulty and practicality' of associating the data with an identifiable individual. This too is multifaceted. For example, with regard to considering a 'reasonable member of the public', does 'reasonableness' bear upon a person's technical or professional expertise? Or upon personal or professional motivations for identifying the individual?[59] The minimal case law addressing this question indicates that

51 See eg, *WL v Randwick City Council (GD)* [2007] NSWADTAP 58 [15].

52 [2015] FCA 1068 (*Baptist Union*).

53 Ibid [49]–[52] (Rangiah J). His Honour referred to the amended definition of 'personal information' in the *Privacy Act* (and the ALRC recommendations upon which it was based) which provided that should a person be identifiable through linkage with information other than the information in question, this would make the information being considered personal information.

54 We note that under the current review of the *Privacy Act*, it has been proposed that 'reasonably identifiable' should be defined to refer to where an individual may be identified directly or indirectly, and that a list of factors supporting this assessment be included (Australian Attorney-General's Department, *Privacy Act Review Discussion Paper* (n 7) 28). If adopted, this proposal could potentially solidify the requirement for an evaluative assessment in statute.

55 *Baptist Union* (n 52) [51], [53] (Rangiah J). His Honour cited the recommendation of the ALRC in Australian Law Reform Commission (n 34) paras 6.55, 6.57.

56 [2017] FCAFC 4 (*Telstra*). *Telstra* involved an appeal from an Administrative Appeals Tribunal (AAT) decision regarding the interpretation of the words 'about an individual' in the definition of personal information in the Privacy Act.

57 *Telstra* ibid [63].

58 OAIC, *Australian Privacy Principles Guidelines, Privacy Act 1988* (n 36) 20–21, paras B.91–B.93; OAIC, *What is Personal Information*? (n 37) 8.

59 See the discussion in *Jonathan Laird and Department of Defence* [2014] AICmr 144 (*Laird*). This case was a decision regarding whether the public interest exemption of personal privacy (*Freedom of Information Act 1982* (Cth) (*FOI Act*) s 47F) precluded grant of a FOI request. *FOI Act* s 4 provides that 'personal information' under that Act has the same meaning as under the *Privacy Act*. The Australian Information Commissioner undertook an assessment of the context and circumstances determining whether the information in question was identifiable, personal information and determined that it was 'impractical for a reasonable member of the public', being overly 'time-consuming and costly'. The factors considered include the resources and expertise available to the person receiving the information (*Laird* [16]–[17]).

identifiability depends on factors including resources and expertise available to the person receiving or accessing the information.[60] In the context of research, the National Statement directs that determining identifiability must include evaluating the 'type and quantity' of the information, as well as 'any other information held by the individual who receives the information and the capacity (skills and technology) available to the individual who receives it'.[61]

The evaluative process in determinations of identifiability confers significant responsibility on the actors involved in collecting and sharing genomic data, requiring them to exercise a degree of discretion in determining if the data are privacy protected. The evaluative process of assessing identifiability is also dynamic and ongoing—for example, the OAIC provides the following guidance on the meaning of 'personal information' in the *Privacy Act*, '[i]nformation holdings can . . . be dynamic, and the character of information can change over time.' This means that determinations of identifiability with regard to the same data may change with new developments and changing environments.[62] Although genomic data are stable, their status with regard to identifiability is dynamic. The determination of identifiability depends on analysis of the genomic data in a particular data situation, at a particular time.

The fluidity of the concept of identifiability is also emphasized in guidance provided in the National Statement for sharing genomic data for research purposes.[63] The National Statement refrains from use of the terms '"identifiable", "potentially identifiable", "re-identifiable", "non-identifiable" or "de-identified" as descriptive categories for data or information due to ambiguities in their meanings. . . .'[64] It describes 'identifiability' as a fluid characteristic existing on a continuum and impacted by context.[65] While the uses of the terminology in the National Statement are not wholly consistent with the guidelines and explanatory material of the *Privacy Act*, the view that identifiability of information is fluid and existing on a continuum is not at odds with the Act.[66] This fluidity of interpretive contexts, and thus of identifiability itself, undoubtedly adds to the complexity and challenges the sufficiency of current laws regulating genomic data-sharing.

In the current review of the *Privacy Act*, consideration is being given to whether the distinction between personal information and 'de-identified' information should be made sharper. It has been proposed under the review that for information to fall outside the definition of 'personal information' and for the Act to no longer apply, it should be classified as 'anonymous'.[67] The aim of such of reform would not be to impose an 'absolute or unworkably high standard' which may impede the use of data for research or health service provision, but to require that only where 'the risk of re-identification was extremely remote or hypothetical' would information fall outside the scope of the Act.[68] In our view, this proposal retains the fluidity of 'identifiability' in the *Privacy Act* and the need for an evaluative assessment, but presents a perhaps welcome clarification that an extremely low risk of (re)identification would be required for genomic data to be shared freely.[69]

The significance of association

Where particular genomic data are placed on the 'identifiability continuum' depends upon the potential for association with other information in the relevant 'data environment'. While genomic data itself is unchanging over time, more data and better linkage techniques are likely to become available, affecting an individual's 'identifiability'. The determination of identifiability

60 *Laird* [16]–[17] ibid.

61 Australian Research Council and Universities Australia (n 13) 34.

62 OAIC, *What is Personal Information*? (n 37) 6, 10. See also the guidance of the OAIC regarding de-identified information, that '[t]he same information may be personal information in one situation, but de-identified information in another' (OAIC, *De-identification and the Privacy Act* (n 50) 14).

63 Australian Research Council and Universities Australia (n 13). The National Statement is not legally binding, but compliance is a prerequisite for funding by the NHMRC, a key public funding body, and the statement in effect plays a role in the 'broader regulatory regime' governing human research, including genomic research.

64 Australian Research Council and Universities Australia (n 13) 33, n 3.

65 Ibid 33–34. The National Statement refers to some additional contextual factors which go to the identifiability of genomic information used for research, including features of a research project such as whether the participant cohort includes high-profile (publicly known) individuals, or whether it involves small communities or large populations.

66 In its submissions regarding review of the *Privacy Act*, the OAIC recommended replacing the term 'de-identified' with 'anonymised', to describe information which is no longer about an identifiable (or reasonably identifiable) individual, and to distinguish between the legal meaning of the term and description of the fluid process of de-identification. OAIC, 'Privacy Act Review – Issues Paper: Submission by the Office of the Australian Information Commissioner' (11 December 2020), 34, para 2.34 <https://www.oaic.gov.au/privacy/the-privacy-act/review-of-the-privacy-act/privacy-act-review-issues-paper-submission/part-2/> accessed 6 August 2022.

67 Australian Attorney-General's Department, *Privacy Act Review Discussion Paper* (n 7) 29–31.

68 Ibid 27,31.

69 The OAIC's submissions to the *Privacy Act* review propose additional protections for anonymised information (OAIC, 'Privacy Act Review – Issues Paper: Submission by the Office of the Australian Information Commissioner' (n 66) 35, para 2.36). The question of whether 'anonymous' data should be privacy protected, and which protections should apply, is beyond the scope of this article. However, it is a relevant and important inquiry with regard to genomic data.

must be made when data are first collected and shared to determine whether *Privacy Act* restrictions apply. However, when genomic data are held longitudinally, it is highly likely that different, and unpredictable, uses of the data, together with technological developments will mean that over time an identified or identifiable individual could be associated with genomic data.[70] This means there is need to periodically review the capability of the relevant data environment to continue to control the potential for association.

Identifiability of genomic data

Data features: uniqueness, volume, and richness

Discussions of the identifiability of different types of genomic data tend to focus on features that impact the risk of identification of data subjects, namely data's uniqueness, and its volume and richness.[71] We consider these in turn.

Uniqueness

'Uniqueness' of genomic data refers to the prevalence of the particular variant(s) or sequence in a population, whether this refers to the general population or a particular sample. A lower prevalence of a particular genetic sequence in the population may be understood to increase the likelihood that the person to whom it relates may be distinguished from others in the population.[72]

The OAIC and CSIRO *Data 61, De-Identification Decision-Making Framework*[73] (*Data 61 Framework*) identifies 'uniqueness' as fundamental in assessing the risk of disclosure of identity.[74] The framework describes data sets containing sets of variables in which the combination of these variables is not shared with any other data set, as 'unique.'[75] Whole genome data sets and larger genomic data sets, which contain more variables are presumed to contain more 'unique' combinations. SNPs that are common to many are considered to represent less risk of being about a particular individual (or potentially—group), if linked to other data, while rare variants, which hold information relevant to few individuals, carry greater privacy risks.[76]

While the distinction between identificatory potential of common and rare variants is acknowledged, we contend that the 'identifiability' of individuals cannot be determined by the 'rareness' of any given variant in abstract isolation. We concur with Wright and others' view that no individual genetic variant is uniquely identifying, not even a rare one.[77] Re-identification of a data subject would 'require an intimate knowledge of the individual's genotype or phenotype together with some information to trace that genotype/phenotype to a specific person'.[78] This is because identifiability requires association, ie, the combination and interaction between the genomic data and associated information in the data environment. This does not denote that the rareness or uniqueness of variant information does not bear upon identifiability, but that the data sharing environment is a critical factor in determining the identifiability of such information.

WGS data, in contrast to variant data, are broadly described by many as unique to the individual and inherently identifiable information.[79] WGS data are sometimes described as 'identifying' or more prone to 're-identification' compared to individual variant information, which is described in the literature as not 'identifying'.[80] Indeed, WGS data have been referred to as an 'ultimate identifier' because each person's whole

70 Finnegan and Hall (n 9) 27.

71 Ibid 11. By 'rich data' we refer to data that are revealing of the 'complexities and the richness of what is being studied'. Sherry Marx, 'Rich Data' in *The SAGE Encyclopedia of Qualitative Research Methods* (2008) 794–95 <https://sk.sagepub.com/reference/research/n408.xml> accessed 6 August 2022. It may refer to 'bigger data' or more 'complex' data. Data complexity may refer to its structure, whether form a single or multiple sources, the quality of the data, the interaction or complexity of relationship between variables, whether it is static or changing. Martin Sheppard, *CS5702: Modern Data Book* (Bookdown 2021) para 2.2 <https://bookdown.org/martin_shepperd/ModernDataBook/C2-Intro.html>

72 Recall that according to the OAIC guidance, information is about an identified person when that person can be 'distinguished' from all others in the group through linking the individual with the information (OAIC, *What is Personal Information*? (n 37) 8).

73 CM O'Keefe and others, *The De-Identification Decision-Making Framework* (CSIRO Reports EP173122 and EP175702, 18 September 2017) <https://www.data61.csiro.au/en/Our-Work/Safety-and-Security/Privacy-Preservation/Deidentification-Decision-Making-Framework> accessed 6 August 2022.

74 Ibid app B, 17.

75 Ibid. The authors wrote, 'A record is unique on a set of key variables if no other record shares its combination of values for those variables.'

76 See eg Sobia Raza and others, *Data Sharing to Support UK Clinical Genetics and Genomics Services*, Workshop Report (PHG Foundation 2015) 27. Azzariti and others suggested that such data should be shared regardless where the risk of identification of an individual was outweighed by the potential harms incurred from not sharing the information, but that supporting data relating to phenotype, or other 'sensitive information' (such as HIV status) or unique information (such as 'an isolated ethnic group) should be kept to a minimum. (Danielle R Azzariti and others, 'Points to Consider for Sharing Variant-Level Information from Clinical Genetic Testing with ClinVar' (2018) 4(1) Molecular Case Studies a002345, 5.)

77 Caroline F Wright and others, 'Genomic Variant Sharing: A Position Statement' (2019) 4(22) Wellcome Open Research, 5 <https://wellcomeopenresearch.org/articles/4-22> accessed 6 August 2022.

78 Ibid.

79 See eg, Raza and others (n 76) 27; Wright and others (n 77) 5, who describe the perception that all genetic data is personal comes from conflating whole genome data with individual genetic variants.

80 See eg, Raza and others (n 76) 27; Finnegan and Hall (n 9) 12; Jean Louis Raisaro, Erman Ayday and Jean-Pierre Hubaux, 'Patient Privacy in the

genome is distinct.[81] Those promoting open and free sharing of variant information, as compared to WGS/WES data, distinguish between WGS/WES data and individual genetic variants, which they propose are not identifiable and pose less of a privacy protection issue.[82] For example, Shabani and others[83] wrote that databases such as ClinVar address privacy concerns by holding only variant data, and not large genomic data sets and that 'multiple variants from the same individual are not linked together in the database'.[84]

Heeney and others discuss a second aspect of 'uniqueness': the population from which the data were taken.[85] Data sets may contain 'special uniques' when they are derived from individuals or units which are unique in the general population. In other words, 'special uniques' are qualities which link the data to an individual's affiliation with a particular category or unit, or could link sample wide genomic data to a group or unit distinguishable from others in the general population. These could arise in units such as families or households; however, they could also occur due to ethnicity or township of residence.[86] Recall that genomic data, whatever its form, holds information that is not only about the person from whom the DNA sample was taken, but also about their family and potentially about groups of which they are a member.[87]

'Uniqueness' in the literature thus refers both to uniqueness within a particular 'population data file' and the uniqueness of a unit within a sample file. Both aspects of 'uniqueness' comprise important elements in assessing identifiability.[88] However, we contend that approaches relying strictly on 'uniqueness' in determining identifiability may be hugely misleading. Both variant data (including rare and common variant data) and WGS may distinguish an individual in the right context.[89] It has been shown that individuals can be re-identified from a range of different forms of genomic data[90] and that individuals can even be re-identified from aggregated whole genome data.[91] For example, Pakstis and others proposed a method for 'uniquely identifying every individual' for purposes such as forensics and paternity testing, using SNPs as 'best markers' for the purpose.[92]

In the Section 'Categories of genomic data', we distinguished between WGS/WES raw sequence read data and interpreted data. One significance of this distinction is that raw data, stored over time, may have greater potential for reanalysis and reinterpretation through previously unavailable methods.[93] This means that possibilities for new findings generated and for linkages are open ended and that identifiability of the genomic data may vary. This may be compared to interpreted genomic variant data.[94] As Azzariti and others discuss, such information pertains to a generalized description of a variant, and not to any individual patient.[95] While all genomic data may be reanalysed and reinterpreted, data which has already been interpreted to a significant level of generalization is less given to further reanalysis and new potential linkages to an identified or reasonably identifiable individual. However, as we have emphasized, this distinction is a matter of degree, and subject to contextual factors.

Volume and richness

The larger volumes of data generated in whole genome sequencing have been tied to 'uniqueness' because the increased possibility of unique combinations of variables. The volume and richness of WGS/WES data also

Genomic Era' (2014) 103(1) Praxis 579, 579. The authors termed 'the genome itself' as the 'ultimate identifier'; William W Lowrance and Francis S Collins, 'Ethics: Identifiability in Genomic Research' (2007) 317(5838) Science 600, 601.

81 See eg, Raisaro and others ibid; Erman Ayday and others, 'Whole Genome Sequencing: Revolutionary Medicine or Privacy Nightmare?' (2015) 48(2) Computer 58, 62.

82 See Wright and others (n 77) 6; Azzariti and others (n 76) 2–3; Mahsa Shabani and others, 'Variant Data Sharing by Clinical Laboratories through Public Databases: Consent, Privacy and Further Contact for Research Policies' (2019) 21(5) Genetics in Medicine 1031, 1034.

83 Shabani and others ibid.

84 See Shabani and others (n 82) 1034.

85 C Heeney and others, 'Assessing the Privacy Risks of Data Sharing in Genomics' (2011) 14(1) Public Health Genomics 17.

86 Ibid 20.

87 See Raza and others (n 76); Heeney and others (n 85), 19. See also O'Keefe and others (n 73) 30–32. A 2008 study demonstrated that sibling identities could be established from published sequence data, and that this could be done on the basis of a 'very low number' of matches of common SNPs, through inferring over half of the sibling's allele frequency data from population specific allele data, and knowledge of genotype of another sibling. See Christopher A Cassa and others, 'My Sister's Keeper?: Genomic Research and the Identifiability of Siblings' (2008) 1(1) BMC Medical Genomics 32.

88 Heeney and others (n 85) 20; O'Keefe and others (n 73) app B, 18.

89 See Finnegan and Hall (n 9) 12.

90 See Laura L Rodriguez and others, 'The Complexities of Genomic Identifiability' (2013) 339 Science 275.

91 See Heeney and others (n 85); Nils Homer and others, 'Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays' (2008) 4(8) PLoS Genetics <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2516199/> accessed 6 August 2022; M Gymrek and others, 'Identifying Personal Genomes by Surname Inference' (2013) 339(6117) Science 321; Rodriguez and others (n 90); see also Caroline F Wright, Matthew E Hurles and Helen V Firth, 'Principle of Proportionality in Genomic Data Sharing' (2016) 17(1) Nature Reviews Genetics 1, 1.

92 Andrew J Pakstis and others, 'SNPs for a Universal Individual Identification Panel' (2010) 127(3) Human Genetics 315.

93 See Mahsa Shabani, Danya Vears and Pascal Borry, 'Raw Genomic Data: Storage, Access, and Sharing' (2018) 34(1) Trends in Genetics 8.

94 See Wright and others (n 77) 6, Table 1. These variables would be considered supporting data in our analysis, discussed further in the following subsection.

95 Azzariti and others (n 76) 3.

creates more opportunities for linkages with associated information from other sources and the possibility of non-deliberate, 'spontaneous recognition'.[96] This also applies to 'larger genomic data sets' including variants identified in large multigene panels.[97] Additionally, those proposing that 'de-identified variant information' may be shared without consent point to the reduced volume and richness of variant data as compared to WGS/WES, or large multigene panel data.[98] In our view, this proposition is overly simplistic. The significance of volume or richness in data depends upon how, in the particular data situation, its volume or richness facilitates linkages with associated information.

We suggest that the strong emphasis on features of genomic data, like its uniqueness and its volume and richness, are misplaced. It can support misplaced inclusion and exclusion. No genomic data are identifying without the right context, but any genomic data can be identifying in the right context. What is more, since the *Privacy Act* definition of personal information was amended, the question of whether an individual's identity can be reasonably ascertained from genomic data has only been one among a number of relevant considerations.[99]

Other aspects of the 'data situation'

The attributes of uniqueness, volume, and richness of genomic data *are* significant—they facilitate linkages between these data and other information associating it with particular individuals. These attributes also reduce the number of individuals with whom the data may be associated, increasing the accuracy of associations drawn. However, they are not the only factors and should not be considered in isolation. The *Data 61 Framework*[100] uses the term 'data situation' to describe the 'relationship between the data and its environment'.[101] Assessing identifiability in the 'data situation' means studying the interaction between the data and the environment in which the data are held, processed, or shared. This evaluation is complex and multifactored and needs to consider genomic data in relation to its supporting data, as well as the broader data environment.

Supporting data

Supporting data attached to genomic data plays an important role in assessing identifiability and determining whether genomic data are personal information. Wright and others distinguished between the types of privacy protection required for genetic variant information based on the supporting data attached to it.[102] For example, they compared individual genetic variants with minimal clinical information attached ('clinical deidentified variant sharing') to variant data with detailed case-specific information'.[103] They recommended that the first category should be open to sharing without patient consent, while the second should remain in the health care system in which it was generated, to be shared openly only with patient consent.[104] We support Wright and others' view that 'different levels of clinical detail will require different modes of sharing, ie, open versus controlled access'.[105] We would add that it is not only the level and volume of evidence but the capacity to cross reference the evidence with auxiliary information which impacts upon identifiability. This assessment of 'identifiability' is equally informed by the features of the data, the attached supporting data, and the mode of access in the data sharing environment.

The *Data 61 Framework* distinguishes between 'direct' and 'indirect' identifiers. 'Direct identifiers' are 'any attributes or combination of attributes that are structurally unique for all persons. . .' in the data.[106] Supporting data such as a person's name, address, or Australian Medicare number, are considered 'direct identifiers'. However, as we have argued, any identifier, even a 'direct identifier' (such as Medicare number) may not be sufficient to relate it to an individual unless it can be linked to associated information and it is 'reasonably practicable' in the circumstances to do so. (For example, a Medicare number acts as an identifier only where the entity accessing the data has access or can obtain access to other information linking an individual to the Medicare number, without unreasonable cost or difficulty.)

The *Data 61 Framework* describes 'indirect identifiers' as data which 'can be used to identify an

96 OAIC, *De-identification and the Privacy Act* (n 50) 9.
97 As described by Azzariti and others (n 76) 3.
98 Wright and others (n 77) 3; Azzariti and others (n 76) 3.
99 See Finnegan and Hall (n 9) 12, 22.
100 O'Keefe and others (n 73).
101 Ibid 25, 66.
102 Wright and others (n 77) 6–7.
103 Ibid.
104 Ibid. See also Shabani and others (n 82) 1034. In practice, there is often some confusion regarding which consents apply to sharing genomic data collected in a clinical setting. Patients may agree to sharing the data when collected in the clinic, but then do not consent to its use on research consent forms. Although the question of appropriate consent frameworks is beyond the scope of this article, it is an important issue in the sharing of genomic data.
105 Wright and others (n 77) 6.
106 O'Keefe and others (n 73) 32.

individual with a high probability, either alone or together with auxiliary information containing other indirect identifiers.[107] It considers indirect identifiers to include variables such as gender, date of birth, country of birth, birth weight, geolocation, language spoken at home, ethnic origin, marital status, total income, dates of medical procedures, diagnoses, etc., dates of hospital admissions, diagnosis, or procedure codes.[108] For example, supporting data attached to annotated variant data are likely to include the genetic condition and its inheritance pattern (if known), the data subject's phenotype (clinical presentation), and a 'cryptic or hidden link' to the submitting entity, to allow further information and follow up.[109] The 'evidence' may also include information available in relevant literature and published analyses, as well as 'a summary of the laboratory's case-level experience with the variant' including the number of persons in which the variant was identified and a summary of the phenotypic and demographic features of those individuals.[110] The *Data 61 Framework* posits that '[a]lmost any variable can be an indirect identifier depending on the auxiliary information available'[111] and suggests that '[w]hen in doubt, it is safest to assume a given variable is an indirect identifier.'[112]

We agree that 'almost any variable' can have identificatory relevance. However, we argue that any data, even those generally considered 'direct identifiers', are only identifying in particular data situations. The distinction between 'direct identifiers' and 'indirect identifiers' is to some extent artificial. However, for our purposes, we acknowledge those variables which routinely function as identifiers and, in relation to which, the means of linkage to an individual are relatively accessible. Thus, for example, variables such as name, address, telephone numbers, email addresses, and license numbers are considered ostensibly 'identifying' and are often removed or modified at the outset when seeking to de-identify information, while other variables with identificatory potential, included as 'supporting data', are shared. We emphasize that, as with all aspects of assessment of identifiability, whether particular supporting data facilitates an association with an identifiable individual is an evaluative question, taking into account the 'likelihood, cost, difficulty and practicality' of linking the data to an individual in the particular data situation.

The data environment

We have emphasized that the whole 'data situation', ie, the relationship between the genomic data in question and the nature of the data environment in which it is held or shared, must be considered in evaluating identifiability. A data situation is dynamic when data are shared between environments or environments change. There may also be multiple data environments, such as occurs when there is a multi-entity data sharing arrangement, or when the data originally shared in one environment are 'on shared' to others.[113] In each one of these changing data situations, the likelihood for association with an identifiable individual varies. As Wright and others wrote,[114] assessing the relationship between the nature of the data and the degree to which access is controlled and restricted is an exercise in proportionality, balancing the 'depth of the data' and the 'breadth of the sharing'.[115]

Ultimately, those collecting, processing and sharing genomic data must evaluate each data situation on its facts. The burden placed upon clinicians and researchers as data controllers is potentially significant. Given current law, perhaps the best way forward is to set out the elements which can guide the assessment. We suggest that the following three data environment elements are useful when assessing genomic data sharing situations: (i) whether the data are shared in a restricted access or open access (public) environment; (ii) the availability and accessibility of other associated data in the relevant (restricted or open) environment; and (iii) who is receiving or accessing the data.

Open or restricted access. Dove and others described 'two main forms' in which genomic data are made available: 'open/restricted access' and 'open/restricted data'.[116] The latter may be created by applying particular processes, eg, de-identification processes, to the data itself. The former relates to access requirements, focusing on the people and the environment to which the data are made available.[117] The level of access flows from what Dove and others term the 'modalities of data sharing' for genomic data, namely: 'ad hoc response[s] to requests' for genomic data, 'collaborations', and 'open access' sharing.[118] Sharing within a collaboration, or in response to a specific request occurs between

107 Ibid.
108 Ibid.
109 Wright and others (n 77) 5.
110 Azzariti and others (n 76) 4.
111 O'Keefe and others (n 73) 32.
112 Ibid.
113 For a description and analysis of various data sharing situations, see O'Keefe and others (n 73) 40–44.
114 Wright, Hurles and Firth (n 91) <https://pubmed.ncbi.nlm.nih.gov/26593419/>
115 Ibid.
116 Edward S Dove, Graeme T Laurie and Bartha M Knoppers, 'Data Sharing and Privacy' in Geoffrey S Ginsburg and Huntington F Willard (eds), *Genomic and Precision Medicine* (Elsevier 2017) 143, 147.
117 Ibid.
118 Ibid.

specified parties within a relationship, is subject to certain agreements, and by its nature will have a more restricted user base. Within a more restricted, controlled environment, the entity sharing the genomic data can use technical, organizational, and legal measures to safeguard the data and control further usages.[119] However, unrestricted, open access of genomic data is often sought for its broad benefit to both research and clinical use.

Public sharing of genomic data significantly limits the sharing entity's ability to monitor who, for what purposes and in what data environments the data will be accessed and what it will be used for. Where data are shared publicly, the 'user base' is 'potentially the whole world'.[120] This increases the risk that someone amongst the data users will hold or have access to information which (when linked with the genomic data) associates it with an identified or identifiable individual. Even if there is no particular motivation to create such linkages, the possibility of 'spontaneous recognition' should also be considered. This refers to the 'unmotivated identification of an individual in a data set from personal knowledge of a small number of characteristics'.[121]

In view of these issues, OAIC guidance recommends that open access environments are only appropriate for data that is either not personal information to begin with, or which has been de-identified so that it is no longer personal information.[122]

Accessibility of associated information. The test of 'reasonable identifiability' does not require that there is no risk of identifiability, but only that it is a very minimal risk. As stated by Rodriguez and others, complete de-identification is unrealistic in today's 'data rich' society and the task at hand is to establish, on a continuum of identifiability, the level of risk posed by particular data in context. We have emphasized that this question is an evaluative assessment and a matter of degree. The amended definition of 'personal information' in the *Privacy Act* broadens the scope of 'personal information', requiring only that the data are about an identified or identifiable individual, and not that the data itself contributes to the identification of the individual. Accordingly, if genomic data are about an individual who is reasonably identifiable, it is captured (and will be 'personal information'), regardless of whether the genomic data are used to actually identify the individual.

The information in the broader data environment is critical. When data are shared publicly, information may be accessed through public registers (such as the electoral register) or social media. Publicly available data includes commercial data bases (eg, commercial lifestyle databases which can be accessed on a user-pays basis), as well as data which can be accessed at little or no cost, by anyone searching for an individual or group. Such data provides information such as: address, age, sex, qualification, workplace, occupation, cultural and ethnic group, ancestry, religion, country of birth, and marital status.[123] There is also 'local knowledge' available to certain segments of the population, such as house details available to real estate agents or student information available to school staff. In certain environments, such as some clinical organizations, associated information about the population which contributed to the genomic data may be available. Additionally, there is the personal information held by people about individuals with whom they have relationships—friends, family, neighbours, and colleagues. Where genomic data are publicly shared, all these types of auxiliary information are available within the 'whole world' of various potential data users.

The ability to identify individuals through linking genomic data with publicly available data, including genealogy data, has been demonstrated in a number of 're-identification' studies.[124] These studies involved different types of genomic data and other data, which had been statistically aggregated or otherwise 'de-identified'.[125] Studies have also illustrated that 'siblingship' could be determined from published SNP sequence data.[126] What this research broadly indicates is that regarding different types of genomic data, and despite de-identification processes, the accessibility of associated information means that sharing genomic data openly presents some degree of risk of identifiability.

In restricted access environments, while controls may be more effectively put in place, there may be additional 'other information' available specifically to persons

119 See generally data sharing scenarios described in OAIC, *De-identification and the Privacy Act* (n 50) 3; O'Keefe and others (n 73) 25–28.

120 See O'Keefe and others (n 73) app B, 13.

121 Ibid app B, 12.

122 O'Keefe and others (n 73) 21. OAIC, *De-identification and the Privacy Act* (n 50); OAIC, *What is Personal Information*? (n 37).

123 See O'Keefe and others (n 73) app E, 46–50.

124 See eg, Rodriguez and others (n 90); Gymrek and others (n 91) 275–76; Homer and others (n 91); Nicholas Masca, Paul R Burton and Nuala A Sheehan, 'Participant Identification in Genetic Association Studies: Improved Methods and Practical Implications' (2011) 40(6) International Journal of Epidemiology 1629; Chris Culnane, Benjamin Rubinstein and Vanessa Teague, 'Health Data in an Open World' [2017] *arXiv:1712.05627* <https://arxiv.org/abs/1712.05627>

125 Gymrek and others (n 91); Masca, Burton and Sheehan ibid; Homer and others (n 91); Culnane, Rubinstein and Teague ibid.

126 See Cassa and others (n 87). The authors demonstrated that sibling genotypes could be inferred using 'proband SNP data' and 'population-specific allele frequency databases' (HapMap).

within an organization accessing the data. In particular, the *Data 61 Framework* identifies a wide range of 'auxiliary information' generally available to a user within organizations dealing with data sets containing health information. Such information includes workplace, employment status, number of dependent children, long-term illnesses, distance of home to work, household tenure, and Australian and New Zealand Standard Classification of Occupation (ANZSCO).[127] Thus, even in a restricted environment, the availability of such auxiliary information must be considered in deciding if genomic data can be shared and what type of controls and safeguards may be necessary.

The data recipient or accessor. According to OAIC guidance, an assessment of 'reasonable identifiability' must also consider the people or organizations who hold and access the information, other information these entities may hold and their capacity to use resources available to them to identify an individual to whom the data relates. Where data are shared publicly, the people or organizations taken into consideration are epitomized in the 'reasonable member of the public'.[128] There is little precedent to guide us regarding the characteristics of the 'reasonable member of the public'; however, the minimal case law available[129] notes factors such as the resources, capacity, and expertise available to the recipient or accessor. Where the user base is potentially the 'whole world', it is difficult to draw parameters around what the 'reasonable' resources, capacity, and expertise available to a data accessor would be. One may also ask what role motivation plays in assessing the likelihood that links will be drawn associating the genomic data with a particular individual. We have already noted that inadvertent or spontaneous recognition is more likely when data are publicly shared, but with regard to intentionally linking genomic data to an individual, the element of motivation is extremely difficult to assess.

Heeney and others suggest applying the concept of the 'data intruder', a person 'with a motivation to investigate the attributes or identity of a data subject, and who uses available information for reidentification of individuals'.[130] They note that the 'motivation to investigate 'may exist for a wide variety of reasons and need not be sinister, but that it is not feasible to anticipate when a reasonable member of the public will be a 'data intruder'.[131] It is therefore necessary to work with the 'likelihood, cost, difficulty and practicability'[132] of identifiability, assuming that a data intruder exists. Where access is restricted and governed by contractual arrangements, restrictions on the use of skill and expertise to draw associations linking the data to an individual may be dealt with in the terms of the agreement. In an open access situation, the capacity and skills of recipients or accessors of the data are unknown, and there is little that can be done to limit and monitor the conduct of a potential 'data intruder'. This contributes to the likelihood of an association with an identifiable individual being made.

Applying controls and safeguards in the data environment

The *Data 61 Framework* describes four ways in which data are released outside of an organization.[133] The first is open access, with no restrictions as to who may access the data or what they may do with it, and usually there is no monitoring or reporting on the use of the data. Even if users are required to agree to the terms and conditions of use limiting the purposes for which the data may be used and disclosed, ensuring compliance would be difficult. Generally, genomic data shared on a completely unrestricted platform, even having undergone de-identification modifications, are at risk of being personal information under the *Privacy Act.*

The framework also refers to 'delivered access' in which a user must apply and agree to the conditions for use of the data, and 'on-site safe settings', which require the user to access the data at a particular secure location, usually subject to governance controls.[134] Lastly, the framework lists 'secured virtual access' as what is 'widely regarded as the future of research data access'.[135] This refers either to access through a remote network interface in which the user can analyse the data, but where output is monitored in a similar way to an 'on-site safe setting'; or, 'analysis servers' in which users can interrogate the data, without being able to view it, and the analysis server returns the requested results after checking them for 'disclosure risk'.[136]

There are many variations to the configurations of a data environment within the four types set out by the *Data 61 Framework*. Factors such as who has access,

127 O'Keefe and others (n 73) app D, 44–45.

128 OAIC, *Australian Privacy Principles Guidelines, Privacy Act 1988* (n 37) 20–21, paras B.91–B.93; OAIC, *What is Personal Information*? (n 37) 8.

129 *Laird* (n 59).

130 Heeney and others (n 85) 20; see also O'Keefe and others (n 73) 43–44.

131 Heeney and others (n 85) 20.

132 See *Baptist Union* (n 52) [51], [53] (Rangiah J); Australian Law Reform Commission (n 34) paras 6.55, 6.57, and discussion earlier at Section 'What Does it Mean for Information to be 'Identifiable'?'.

133 O'Keefe and others (n 73) app C, para C.2.3.

134 Ibid.

135 Ibid.

136 Ibid.

where they have access (eg, whether within on-site safe setting), data security, data use agreements, and monitoring and reporting protocols all go towards the assessment of whether the genomic data in question, having undergone modifications, may be dealt with as 'de-identified information'. We emphasize that whether modified genomic data are sufficiently de-identified, depends upon whether in the relevant access conditions, it is still possible to link the data with other information associating them with an identified or identifiable individual.

Conclusions

This article began with the question: when are genomic data 'personal information' and (as such) subject to Australian privacy regulation? To address this question, this article has considered different categories of genomic data, the definitions of 'personal information' and 'genetic information' in the *Privacy Act*, and the legal test of 'reasonable identifiability' as applied to genomic data. The analysis has shown that a single, formulaic answer to the question of when genomic data are personal information is not possible. Rather, the article clarifies the factors and considerations relevant to the determination of the reasonable identifiability of genomic data, such as those that would be applied by a court.

A key finding of this article is that the legal and scientific concepts of genetic information (or genomic data) have *separate epistemic bases*. As such, the labels applied to observed data relating to an individual's genetic or genomic makeup may be inconsistent across these two domains. The *scientific* conception of genomic data is associated with distinct methods of discovery and implies knowledge of particular biological features. The *legal* conception of genetic information is a sub-category of, and is therefore dependent upon, the higher order legal concept of 'personal information' in Australian privacy law: If genomic data are not personal information, then they cannot be genetic information. When applying the law to what in scientific terms are 'genomic data', one may enquire whether the boundaries of the distinct concepts are coterminous. In other words, when applying the law to what in scientific terms are 'genomic data', one should enquire whether the legal and scientific concepts align. Therefore, for the sake of analysis and effective communication, it may be useful to use the legal term 'genetic information' to refer to the conception recognized by law and to reserve use of the term 'genomic data' to that implied by scientific, and lay, usage.

'Personal information' in the *Privacy Act* must be about an identified or identifiable person. The literature pertaining to the identifiability of genomic data has focused on the characteristics of uniqueness, volume, and richness. Yet while uniqueness, volume, and richness are relevant considerations, they cannot, when abstracted from a particular data environment, be considered determinative. Identifiability is assessed with reference to the overall 'data situation'—the data, the context, and the relationship between them. An assessment of the overall 'genomic data situation' includes examining the features of the genomic data in question (in conjunction with any supporting data attached) *and* contextual factors such as the degree to which access to the data is free, open, controlled, or restricted, the accessibility of associated information in the data environment and the 'data recipients'—meaning the people or entities accessing or receiving the genomic data. It is only through examining each and all of these factors, and the manner in which they interact, that the identifiability of a person the genomic data are 'about' can be evaluated and the question of whether the data are personal information answered. A data situation may also be dynamic (ie, is not 'set and forget'), shifting as a result of technological developments or in multiple or changing processing environments. The evaluation of identifiability is therefore an ongoing process, and the status of genomic data may change over time.

Setting out a practical step-by-step framework for assessing the identifiability and the legal status of genomic data *vis-à-vis Privacy Act* protection is beyond the scope of this article. However, the analysis and clarification provided point to the processes which must be addressed in such guidance. These include: examination of the relevant features of the genomic data, and in particular uniqueness, volume, and richness; consideration of the presentation of the data and whether and what kind of supporting data are attached to it; assessment of the data storage and data processing environment with regard to the type of access it provides; the associated information available with it, the knowledge and skills of those accessing or receiving the data and the controls and safeguards applied in the environment; and, a built-in mechanism for reassessment in accordance with the dynamic quality of the particular data situation. There is no question that such an evaluation is detailed and complex. However, a clear understanding of the factors contributing to the identifiability of genomic data will aid a system of evaluation and ultimately enable the flow of genomic data and protect the privacy of those who contributed them.

Acknowledgement

Draft versions of this paper were presented at the Australian Genomics/Google *Reimagining Health*

Genomics: Technology Summit, the Health Law and Ethics Network at the University of Melbourne and at an Australian BioCommons seminar. The authors thank attendees at these meetings for their helpful feedback. The authors also thank Norah Grewal and Yael Prawer for their assistance and advice.

https://doi.org/10.1093/idpl/ipad002
Advance Access Publication 1 February 2023