

SCIENCE AND SOCIETY

From genetic privacy to open consent

Jeantine E. Lunshof, Ruth Chadwick, Daniel B. Vorhaus and George M. Church

Abstract | Recent advances in high-throughput genomic technologies are showing concrete results in the form of an increasing number of genome-wide association studies and in the publication of comprehensive individual genome–phenome data sets. As a consequence of this flood of information the established concepts of research ethics are stretched to their limits, and issues of privacy, confidentiality and consent for research are being re-examined. Here, we show the feasibility of the co-development of scientific innovation and ethics, using the open-consent framework that was implemented in the Personal Genome Project as an example.

Current developments in genomics challenge the established framework of biomedical ethics because the empirical facts of the genomic science change too fast for the reflections of ethics to keep pace with. At the same time, as practical applications of new technologies are being developed, scientists call for pragmatic moral guidance¹. Recent revelations about the human genome, such as the abundance of copy-number variation (CNV)², and the large-scale identification of functional elements through the [Encyclopedia of DNA Elements \(ENCODE\) project](#)³ pave the way to a new understanding of human genome function. The number of published genome-wide association studies (GWAS) continues to rise quickly. Newly developed technologies, in particular high-throughput, low-cost sequencing^{4,5}, are being applied to increasingly large human genome and phenome data sets. These developments have ethical, legal and social implications that call for strong cooperation between science and humanities⁶. While looking for approaches that can adequately address the moral and policy issues that are raised by emerging genomic technologies, ethicists are increasingly aware of the need for a shift in emphasis, even if it ultimately requires revision of key concepts in mainstream biomedical research ethics.

One component of traditional medical ethics — the obligation to confidentiality — has recently come under review. In addition to its implementation in the clinical setting and in the context of public health, the applicability of confidentiality to large-scale genomic research now calls for attention. Developments in both medical informatics and bioinformatics show that the guarantee of absolute privacy and confidentiality is

not a promise that medical and scientific researchers can deliver any longer^{7,8}. This has concrete implications for the validity of consent for participation in research. Consent is relevant in building large-scale databases containing genotype data that are inevitably traceable to individuals, as well as in generating ‘personal genomes’. What pragmatic moral guidance can be offered under these new circumstances?

In this article we argue that the reality of the new genetics and genomics urges us to abandon the traditional concept of medical confidentiality. As we hold the view that ethical thinking evolves alongside science⁹, we argue that new models are needed to offer robust moral guidance while keeping the reality of a dynamic science in mind. One such new model is the open-consent approach, developed in the [Personal Genome Project \(PGP\)](#). We take this example to illustrate the feasibility of the co-development of ethics and genomics in a specific study protocol. We focus on the quality of consent to participation in studies using correlated genotype and phenotype data.

“...new models are needed to offer robust moral guidance while keeping the reality of a dynamic science in mind.”

Genetic privacy

The emergence of genetic privacy. Genetic privacy usually refers to informational privacy¹⁰. It indicates an individual’s right — one that is perhaps extended to families and communities — to protection from non-voluntary disclosure of genetic information.

This concept emerged over the last few decades, as a consequence of developments in genetics and information technology. What made the rise of this concept possible was the disclosure of the ‘invisible’ part of heredity at the molecular level, prior to which the information about hereditary traits was limited to what could, in principle, be known to others — such as individual and family health history (even if certain diseases running in the family were kept as a family secret), pedigree information and obvious physical traits. More recently, rapid advances in sequencing technologies are making fast and affordable whole-genome sequencing readily available, and developments are continuing to accelerate^{11,12,13}. Comprehensive data sets to establish informatics links among ten thousand to a million human genome sequences and extensive phenotype analyses are needed to effectively generate and test hypotheses, but they also enable the identification of the individuals whose DNA sequences they contain. This puts the validity of the existing consent protocols into question. If promises of privacy and confidentiality need to be abandoned, what are the implications for meaningful consent in the context of genomics research?

Genetic privacy can be taken to denote a particular instance of the general concept of privacy, although often it is used as a value-laden concept that is qualitatively different from ‘normal’ privacy¹⁴ — a concept that presupposes adherence to genetic exceptionalism¹⁵. By contrast, we subscribe to the view that genetic privacy is just one instance of privacy.

Before turning to the general notions of privacy and confidentiality and their relevance to the process of consent, we describe the increasing inadequacy of health information protection against the background of the historical development of information technology and its application for epidemiological purposes.

Health information privacy. The new methods of data storage that were introduced to many hospitals in the 1960s and early 1970s allowed electronic searching and linking¹⁶. This, combined with a growing emphasis on individual rights in medical ethics and health law, highlighted the need for the protection of ‘health information privacy’¹⁷. Since then, the increasing opportunities to gather genetic information about individuals have raised concerns among the public and the policy makers about access to this type of information and its potential abuse. The experience with sickle-cell anaemia screening in the United States demonstrated,

as early as 1972, that stigmatization of individuals on the basis of their membership of a particular group is a real risk¹⁸. Whether this is based on genetic or other traits, conventional individual privacy protection misses the point. It does not work in the case of so-called non-distributive generalizations about groups in which the individual profile is indiscernible from the group profile, as is the case in epidemiological research. The concept of 'categorical privacy' has recently been proposed to overcome the inadequacies of traditional individual-centred concepts of privacy with regard to the individuals that make up the non-distributive profile of a group¹⁹. Current legislative efforts, such as the [Genetic Information Nondiscrimination Act](#) in the United States, attempt to provide a certain level of protection, at least against the potential detrimental use of genetic data²⁰.

Privacy and confidentiality

Privacy is a complex notion. Genetic privacy refers to a specific field of application and is mostly used in the limited sense of informational privacy.

Informational privacy is concerned with the limits on access to personal information; confidentiality, anonymity and secrecy are branches of it¹⁰. Confidentiality implies trust in private and in professional relationships between individuals. The maintenance of confidentiality by professionals is vital to the trust in the profession, for example, to the public trust in physicians, lawyers or members of the clergy. Anonymity refers to a state of blocked or restricted access to information that identifies persons. Secrecy implies having control over the disclosure of information. It entails an aspect of intentional concealment and can also be deliberately used to the detriment of others¹⁰.

Infringement of privacy. Privacy can be violated by forces that are beyond individual or institutional control, such as accidental data release, data release that is required by authorities, or by criminal offences, including burglary, hacking, hardware and/or data theft (BOX 1). Infringement of privacy can cause considerable material and immaterial harm: to social position and opportunities, to personal and familial status, and to self-image and perception by others. However, infringement of privacy need not relate to any moral failure.

In addition to the above threats to secrecy (BOX 1), there is increasing evidence from the medical and bioinformatics fields that indicates that absolute privacy

Box 1 | Threats to privacy and confidentiality

Actions that are aimed at uncovering identity

- Re-identification after de-identification using publicly available data, for instance, finding health records using publicly accessible administrative data (see REF. 58 for an example).
- Combination of surnames as well as genotype and geographical information, for instance, the tracing of an anonymous sperm donor by his offspring (see REF. 59 for an example).
- Inferring phenotype from genotype by identifying information in DNA and RNA, for instance, stature, hair or iris colour, or skin colour (see REF. 60 for an example).
- Any amount of DNA data in the public domain with a name allows for identification within any anonymized data set.
- Identification through the DNA of a first-degree relative, for example, the identification of Bernardo Provenzano through his brother's DNA⁵⁴.
- Identification by phenotype using imaging techniques for reconstruction of facial features.
- Hacking into computer systems.
- Physical attacks on encryption keys; for example, so-called cold boot attacks (see REF. 61 for an example).
- Theft or loss — by accident or forgetfulness — of a laptop or of data-storage devices.

Causes of disclosure of information content

- The increasing availability of aggregate data in public, private and state-controlled databases, including: clinical biobanks and databases; population biobanks and databases; research biobanks and databases with academia and industry; and forensic biobanks and databases.
- Data sharing and secondary use of data.
- Developments in technology, in medical informatics and in bioinformatics.
- Information technology accidents leading to security breaches.
- Actions driven by insatiable curiosity about self and others.
- The increased ease of finding electronic data with web-based search engines.

and confidentiality is not a promise that medical and scientific researchers can deliver^{7,8,21,22}. Malin and Sweeney have shown that re-identification of individuals is possible through genotype-phenotype inference, and through methods such as genealogical information, trail re-identification or so-called dictionary attacks²¹. A lot of effort is made to improve data safety. Recently, the statistic strategy of *k*-anonymization has been developed in which each relevant entity is hidden in at least *k* peers²³. This strategy is used with strong reference to traditional confidentiality in, for example, IBM's Hippocratic Database Technology²⁴. However, it has already been challenged by so-called *L*-diversity, which, according to its proponents, is more robust²⁵.

Breaching confidentiality. A breach of confidentiality implies an action on the part of those who are supposed to keep it. Therefore, in contrast to an infringement of privacy, it implies a moral failing by definition. Maintaining confidentiality, on the other hand, might protect one party from harm while exposing others to it. The paradigm case in medical ethics is the Tarasoff case²⁶, in which a psychiatrist failed to breach confidentiality to warn

a young woman that one of his patients intended to kill her — this woman was subsequently killed by his patient. A 'duty to warn' persons who are directly at risk was derived from this case. Although there are dissenting opinions²⁷, it is widely agreed that the obligation of confidentiality cannot be absolute^{28–30}. The World Medical Association adopted this point of view in its 2006 version of the International Code of Medical Ethics³¹. However, there is also a consensus that specific and weighty circumstances are required to justify a breach of confidentiality. In situations in which there is no threat to the life or well-being of third parties, patients are likely to expect strict confidentiality from their doctors. Therefore, when individuals donate samples and data originating from individual medical treatment for research, they might have unrealistic expectations about the degree of confidentiality that will be provided. This belief is reinforced through reassuring statements about 'strict confidence', contained, for example, in informational materials for research³².

Recent research has shown that in regular general practice, there is also a considerable discrepancy between patients' expectations and the true extent of confidentiality available^{33,34}.

Confidentiality, consent and disclosure.

Awareness of the discrepancy between patients' expectations of confidentiality and actual practice is crucial when devising consent for research participation. These expectations are shaped according to the traditional image of exclusive patient–physician confidentiality. Thus, the perceived confidentiality of the setting in which patient information is generated is decisive. In addition, as Rothstein shows³⁵, there are three key time points with respect to confidentiality. First, the initial moment of sharing of confidential information; in the health-care setting, this is a direct consequence of the patient's decision to seek help. Second, the external disclosure beyond the confidentiality-based relationship when making data available for the purpose of research; explicit consent is needed. Third, the time of potential re-disclosure, for example, through data sharing or linking of data collections in the course of research, or when data that were coded at submission are re-identified. A crucial consideration is that consent for disclosure and re-disclosure is given only upon certain conditions; a key condition usually being the assurance of secrecy with regard to personal identity and information content. Yet, how can secrecy be promised when the sharing of data is not only foreseeable but is, in fact, intended? The promise of secrecy is a major part of our argument in support of the open-consent protocol, as introduced in the PGP.

Consent and re-consent. Both the [UK Biobank](#) and the initial [International HapMap Project](#) are examples of *de novo* data collections and, as such, offer the

unique opportunity to clearly define terms and conditions of consent from the outset. Consent can be narrow and specified, broad, or blanket; blanket consent implies that there are no restrictions to the scope and duration of the consent³⁶. Obviously, broad or blanket consent can never be fully informed³⁷. Consent might include a further layer: the consent to be re-contacted and give re-consent, for example, when new information becomes available that is relevant to the research subject, or if further research is being considered³⁸. However, including the option of re-contacting and obtaining re-consent implies, by definition, maintaining identifiability and traceability of research participants.

In February 2007 the US health-care provider Kaiser Permanente announced a [Research Program on Genes, Environment and Health \(RPGEH\)](#)³⁹ that collects data and samples for GWAS and will surpass the UK Biobank in size and scope. Participants are informed that linking of databases and data sharing among researchers is intended. A striking feature of this project is that, as a health-care insurer, Kaiser Permanente recruits the participants from amongst its members, that is, its own insurees, and uses the data that have been stored in its archives for almost four decades. This is a typical case of new research on extant data for which re-consent will be sought. A preliminary survey, performed by the RPGEH, showed a great willingness to participate. Remarkably, this is in spite of the fact that a serious breach of confidentiality occurred at Kaiser Permanente in 2004, caused by a complex accident in its information technology structures⁴⁰.

False promises, wrong expectations. The language for consent to participation in large-scale studies that require the collection of genotype and phenotype data hardly differs from the traditional consent protocols used in the clinical setting: it emphasizes the protection of privacy and confidentiality. For example, the consent form of the International HapMap Project assures the participants in the following way: "... it will be very hard for anyone to learn anything about you personally from any of this research because none of the samples, the database, or the HapMap will include your name or any other information that could identify you or your family."⁴¹

The HapMap informed-consent protocol does not unambiguously guarantee anonymity or confidentiality of participants' genetic information. On the contrary, it even mentions the risk of tracing identity through publicly available HapMap data. Nevertheless, the consent protocol clearly suggests that the risk of re-identification is vanishingly small.

The information leaflet of the UK Biobank explains to the volunteers that: "Only if you ... give your written consent would we be able to access your medical records. (All such information would be kept in strict confidence)."³²

The data from the available literature are not unequivocal about how many research subjects would withhold consent without the explicit or implicit promise of confidentiality and anonymity. A large survey by Canadian researchers revealed that patients want to be actively consulted and give consent for their personal information to be used for research. The patients, however, make little distinction between identifiable and non-identifiable information⁴². Yet, other studies found that a substantial number of patients do not consent to data collection for research purposes from their existing records^{43,44}.

Non-valid consent. The finding that the confidentiality of genetic data cannot be guaranteed suggests that a research participant's consent might not be valid when it is conditioned on the assurance or even the unchallenged expectation of full genetic secrecy.

At the same time, large correlated sets of comprehensive genetic information are needed for GWAS that aim to uncover genetic determinants of common, complex human disorders^{45,46}. Understanding of biological processes requires integration of diverse types of data. Applying systems-biology

Box 2 | Key features of the Personal Genome Project

The Personal Genome Project aims to build a framework for the development and evaluation of personal genomics technologies and practices at increasing scales. Its key feature is the comprehensive approach towards:

- The development of a broad vision of how personal genomes can be used to improve the understanding and management of disease.
- The development of technologies to improve the affordability of personal genome sequencing.
- The development of tools for interpreting genomic information and correlating it with the individual medical and biological information.
- The development of educational and informational resources for improving general understanding of personal genomics and its potential.
- The fostering of dialogue with research communities, industries, and public and governmental bodies involved with personal genomics and the related ethical, legal, and social issues.
- The development of a normative framework that addresses the needs of personal genome research in the context of open access to comprehensive identifiable genetic information.
- The implementation of this approach in interactive research on human subjects involving individuals who consent to obtaining and openly sharing their genome sequences and their related phenotype information.

The above points are adapted from the [mission statement of the Personal Genome Project](#).

Box 3 | Key features of the Personal Genome Project's open-consent policy

Open consent as part of the Personal Genome Project implies that research participants accept that:

- Their data could be included in an open-access public database.
- No guarantees are given regarding anonymity, privacy and confidentiality.
- Participation involves a certain risk of harm to themselves and their relatives.
- Participation does not benefit the participants in any tangible way.
- Compliance with monitoring of their well-being through quarterly questionnaires is required.
- Withdrawal from the study is possible at any time.
- Complete removal of data that have been available in the public domain may not be possible.

The moral goal of open consent is to obtain valid consent by effectuating veracity as a precondition for valid consent and effectuating voluntariness through strict eligibility criteria, as a precondition for substantial informed consent.

approaches to integrated personal data sets will facilitate the development of new modes of individually targeted treatments or disease prevention^{47,48}. However, anything approaching a comprehensive genotype or phenotype (including molecular phenotypes) ultimately reveals subjects' identities as surely as conventional identifiers such as a name and social security number would. The *American Society of Human Genetics* (ASHG) declares the following in a statement on genome-wide association studies: "[the ASHG is] acutely aware that the most accurate individual identifier is the DNA sequence itself or its surrogate here, genotypes across the genome. It is clear that these available genotypes alone, available on tens to hundreds of thousands of individuals in the repository, are more accurate identifiers than demographic variables alone; the combination is an accurate and unique identifier."⁴⁹

These facts fundamentally challenge current consent practices, ones that strongly suggest or even assure strict confidentiality, in otherwise carefully designed genetics and genomics projects, as illustrated by the HapMap and the UK Biobank. Thus, when applied to GWAS, common and widely used consent practices might in fact result in disingenuous consent, at least insofar as they are based on untenable promises of privacy and confidentiality⁵⁰.

Personal genomes, open consent

How can full or at least substantial^{51,52} informed consent for participation in GWAS be realized? What are the requirements of a study design and a consent protocol that abandon confidentiality in order to preserve trust? Should veracity precede autonomy?

We believe that the building of any comprehensive genotype–phenotype data collection requires that the individuals

from whom these data are derived be fully aware that the data can be and likely will be accessed, shared and linked to other sets of information, and that the full purpose and the extent of further usage cannot be foreseen. Individuals should realize that they are potentially identifiable and that their privacy cannot be guaranteed. Full and valid consent by the participants requires veracity on the part of the researchers, as a primary moral obligation. Below we describe an open-consent model and its practical application within the PGP (see BOX 2 for additional information about the PGP).

The origins of the PGP. The research group that prepared the 2003 National Institutes of Health Center for Excellence in Genome Science proposal for a Molecular and Genomic Imaging Center to the National Human Genome Research Institute (NHGRI), a proposal that aimed to develop ultra-low cost and high-accuracy genomics⁵³, recognized the inadequacy of existing consent practices in the face of the increasing availability of large data sets containing comprehensive identifying genetic information. Therefore, alongside the technology change, a similarly innovative approach was implemented to obtain fully consented comprehensive genetic data sets: the PGP's open-consent protocol (summarized in BOX 3). Transparency is the hallmark of this project, which uses open-source technology and bioinformatics, relies on interactive participation by research subjects and provides open access to data sets that have been consented accordingly.

Open consent. Open consent means that volunteers consent to unrestricted re-disclosure of data originating from a confidential relationship, namely their health records, and to unrestricted disclosure of

information that emerges from any future research on their genotype–phenotype data set, the information content of which cannot be predicted. No promises of anonymity, privacy or confidentiality are made. The leading moral principle is veracity — telling the truth — which should precede autonomy. Although, in clinical medicine, veracity is the legal norm in many jurisdictions, physicians may try to justify the withholding of information by invoking the 'therapeutic privilege'. In research, there is no such privilege, and when seeking informed consent from research subjects, distorted or incomplete information could undermine trust in researchers and in science.

Consenting to disclosure. Whether fully informed consent is just an ideal that cannot have a meaningful place in the practical world⁵¹ and substantially autonomous consent is the best that is attainable is a matter of debate in ethical theory. In the PGP we strive to ensure that the consent process is as fully informed as possible, and results in a substantially autonomous consent. Therefore, the participants of the first 2007 study cohort were requested by the Harvard Medical School (HMS) Institutional Review Board (IRB) to have a master's degree in genetics or equivalent, and have been presented from the outset with a straightforward description of the risks of participation and the harm they might experience as a consequence of the loss of privacy through public disclosure or identification.

Glossary**Genetic exceptionalism**

The view that being genetic makes information, traits and properties qualitatively different and deserving of exceptional consideration.

Non-distributive generalization

Generalizations that entail information about individuals as belonging to a particular group with specific properties. Any particular individual, however, may or may not have these properties.

Dictionary attack

A technique for breaking a security system by trying to determine a decryption key or a password by searching a large number of possibilities.

L-diversity

A new method for the protection of privacy against adversaries with background knowledge, which requires that the distribution of a sensitive attribute in each equivalence class has the least well-represented values.

Open-source technology

A technology that is publicly available, freely distributed by the developer community and that is for the user community to modify and improve.

The envisaged upscaling of the PGP will be guided by the outcomes of the careful monitoring of this initial cohort and by the evaluation of the participants' experiences through their continuous interaction with the project team. Interactive online education and an entrance test will be in place in order to obtain valid consent once the participation is open to the broader public.

The genetic and medical information that is posted on the study website, although it is directly associated only with the research subjects themselves, could also have relevance to participants' family members. Individuals could be traced and identified by any DNA-containing sample from their relatives who might not even be aware of its storage and its possible implications⁵⁴. Although no consent from family members is required by the HMS IRB, in the PGP potential volunteers are strongly advised to discuss their participation with relatives.

Volunteers can withdraw from participation at any time and they can redact specific items in their records at any point in the study. They should be well aware of the fact that items that have been available in the public domain and used, for example, to support conclusions in published work cannot be easily reversed (BOX 3).

It is not promised that participation in the study will benefit the volunteers in any material way. Although individual reasons for participation may remain to a certain extent obscure, and consent might not always be based on purely rational considerations, a high degree of 'information altruism'⁵⁵ is required, thereby introducing a strong moral motive.

Concluding remarks

Current developments in genomic technology challenge the traditional normative framework for biomedical research and its well-known components. It has become clear that the common interpretation of the concepts of privacy and confidentiality as being absolute or near absolute cannot be sustained. Whenever genetic samples are involved re-identification will be possible. Although the research community is well aware of the facts, until now this awareness has not been reflected in the language of consent. Therefore, in many cases, existing consent cannot be assumed to be fully valid.

GWAS are rapidly being implemented. The first results of the [NCBI Database of Genotype and Phenotype](#) are available and are in part publicly accessible⁵⁶. Many more studies that make use of comprehensive genotype-phenotype data are underway, and data sharing in the context of large networks

is an essential part of the research process. In many cases, extant samples and data are being used in a different context and on different conditions from the ones under which they had been collected⁵⁷. This raises serious questions about current consent practices. The burden of proof concerning ethical integrity in the conduct of research with human subjects rests with the researchers. Oversight by ethics committees or IRB approval is no substitute for personal responsibility. An open-minded reappraisal of the relationship between scientists and their research subjects is urgently needed.

New prospective studies provide the opportunity for applying newly devised consent protocols. Here, we have presented the novel open-consent model that has been devised for the PGP — it opts for openness in its scientific design and for veracity as the leading principle in obtaining participant consent.

Alternative solutions are scarce. Veracity requires broad consent in any case of collection and long-term storage of comprehensive data sets. However, an overly broad consent could become meaningless. The most likely pragmatic solution would entail maximizing data protection while informing people about its limits. Proposed solutions to the question of actual ownership of donated data and samples and of intellectual property do not bear upon the issue of promises of anonymity and confidentiality. However, taking into account the trend towards open access, the issues of ownership and benefit sharing will soon call for practical and up-to-date solutions⁵⁷.

Regulation of biomedical research will need to be revised, both at the national and the global level. We believe that sustainable solutions will only be reached through the co-development of the sciences and humanities. The role of ethics is neither that of an alibi nor of a straightjacket.

Jeanine E. Lunshof is at the Department of Molecular Cell Physiology, c/o room M236, Faculty of Earth and Life Sciences, VU University Amsterdam, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands.

Ruth Chadwick is at Cardiff Law School, Cardiff University, Law Building, Museum Avenue, Cardiff, CF10 3AX, Wales, UK.

Daniel B. Vorhaus is at Robinson, Bradshaw & Hinson P.A., 101 North Tryon Street, Suite 1900, Charlotte, North Carolina 28246-1900, USA.

George M. Church is at the Department of Genetics, Harvard Medical School, New Research Building, Room 238, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA.

*Correspondence to J.E.L
e-mail: jeantine.lunshof@falw.vu.nl*

doi:10.1038/nrg2360
Published online 1 April 2008

1. Anonymous. Defining a new bioethic. *Nature Genet.* **28**, 297–298 (2001).
2. Feuk, L., Carson A. R. & Scherer S. W. Structural variation in the human genome. *Nature Rev. Genet.* **7**, 85–97 (2006).
3. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
4. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
5. Shaffer, C. Next-generation sequencing outpaces expectations. *Nature Biotechnol.* **25**, 149 (2007).
6. Daar, A. S., Scherer, S. W. & Hegele, R. A. Implications of copy-number variation in the human genome: a time for questions. *Nature Rev. Genet.* **7**, 414 (2006).
7. Lowrance, W. W. & Collins, F. S. Identifiability in genomic research. *Science* **317**, 600–602 (2007).
8. McGuire, A. L. & Gibbs, R. A. No longer de-identified. *Science* **312**, 370–371 (2006).
9. Knoppers, B. M. & Chadwick, R. Human genetic research: emerging trends in ethics. *Nature Rev. Genet.* **6**, 75–79 (2005).
10. Allen, A. L. in *Genetic Secrets: Protecting Privacy and Confidentiality in the Genetics Era* (ed. Rothstein, M. A.) 31–59 (Yale University Press, New Haven, 1997).
11. Shendure, J., Mitra, R. D., Varma, C. & Church, G. M. Advanced sequencing technologies: methods and goals. *Nature Rev. Genet.* **5**, 335–344 (2004).
12. Check, E. Faster still and faster. *Nature* **448**, 10–11 (2007).
13. Porreca G. J. *et al.* Multiplex amplification of large sets of human exons. *Nature Methods* **4**, 931–936 (2007).
14. Annas G. J., Glantz L. H., Roche P. A. Drafting the genetic privacy act: science, policy and practical considerations. *J. Law Med. Ethics* **23**, 360–366 (1995).
15. Murray T. H. in *Genetic Secrets: Protecting Privacy and Confidentiality in the Genetics Era* (ed. Rothstein, M. A.) 60–73 (Yale University Press, New Haven, 1997).
16. Anonymous. Microfiche system saves time, cuts storage by 98 per cent. *Mod. Hosp.* **107**, 66–67 (1966).
17. Curran, W. J., Hyg., S. M., Stearns, B. & Kaplan, H. Privacy, confidentiality and other legal considerations in the establishment of a centralized health-data system. *N. Engl. J. Med.* **281**, 241–248 (1968).
18. Anonymous. Sickle cell – point, counterpoint. *N. Engl. J. Med.* **289**, 323 (1973).
19. Vedder, A. KDD: The challenge to individualism. *Ethics Inf. Technol.* **1**, 275–281 (1999).
20. National Human Genome Research Institute. Genetic Information Nondiscrimination Act of 2007. *National Human Genome Research Institute [online]*, <<http://www.genome.gov/24519851>> (2007).
21. Malin, B. A. & Sweeney L. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *J. Biomed. Inform.* **37**, 179–192 (2004).
22. Lin, Z., Owen, A. B. & Altman, R. B. genomic research and human subject privacy. *Science* **305**, 183 (2004).
23. Sweeney, L. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **10**, 571–588 (2002).
24. Agrawal, R. & Johnson, C. Securing electronic health records without impeding the flow of information. *Int. J. Med. Inform.* **76**, 471–479 (2007).
25. Machanavajjhala, A., Kifer, D., Gehrke, J. & Venkatasubramanian, M. L-diversity: privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **1**, Issue 1, Article 3 [online], <<http://doi.acm.org/10.1145/1217299.1217302>> (2007).
26. Supreme Court of California, Tarasoff v. Regents of the University of California (17 Cal.3d 425; 551 P.2d 334; 131 Cal. Rptr. 14). *Public Health Law [online]*, <<http://www.publichealthlaw.net/Reader/docs/Tarasoff.pdf>> (1976).
27. Kipnis, K. A defense of unqualified medical confidentiality. *Am. J. Bioeth.* **6**, 7–18 (2006).
28. Beauchamp, T. L. & Childress, J. F. *Principles of Biomedical Ethics* 5th edn 303–312 (Oxford University Press, New York, 2001).
29. Hodge, J. C. The legal and ethical fiction of 'pure' confidentiality. *Am. J. Bioeth.* **6**, 21–22 (2006).

30. EuroCOCAP Project. European Standards on Confidentiality and Privacy in Health Care. *EuroSOCAP* [online]. < <http://www.eurosocap.org/Downloads/European-Standards-on-Confidentiality-and-Privacy-in-Healthcare.pdf> > (2006).
31. The World Medical Association. World Medical Association International Code of Ethics (London, England, 1949). Adopted by the General Assembly (Pilarneberg, South Africa, 2006) [online]. < <http://www.wma.net/e/policy/c8.htm> > (2006).
32. UK Biobank. UK Biobank Information Leaflet (vs. 27.02.2007) *UK Biobank* [online]. < http://www.ukbiobank.ac.uk/docs/InfoleafletEnglish141107all_Pages.pdf > (2007).
33. Carman, D. & Britten, N. Confidentiality of medical records: the patient's perspective. *Br. J. Gen. Pract.* **45**, 485–488 (1995).
34. Stone, M. A., Redsell, S. A., Ling, J. T. & Hay, A. D. Sharing patient data: competing demands of privacy, trust and research in primary care. *Br. J. Gen. Pract.* **55**, 783–789 (2005).
35. Rothstein, M. A. in *Genetic Secrets: Protecting Privacy and Confidentiality in the Genomics Era*. (ed. Rothstein, M. A.) 451–495 (Yale University Press, New Haven, 1997).
36. Hansson, M. G., Dillner, J., Bartram, C. R., Carlson, J. A. & Helgesson G. Should donors be allowed to give broad consent to future biobank research? *Lancet Oncol.* **7**, 266–269 (2006).
37. Árnason, V. Coding and consent moral challenges of the database project in Iceland. *Bioethics* **18**, 27–49 (2004).
38. Dal-Ré, R., Avendaño, C., Gil-Aguado, A., Gracia, D. & Caplan, A. L. When should re-consent of subjects participating in a clinical trial be requested? A case-oriented algorithm to assist in the decision-making process. *Clin. Pharmacol. Ther.* 19 Sept 2007 (doi:10.1038/sj.clpt.6100357).
39. Kaiser Permanente News Center. Kaiser Permanente Unveils Groundbreaking Genetic Research Program. *Kaiser Permanente* [online]. < <http://xnet.kp.org/newscenter/stories/nat/2007-02-15-f.html> > (2007).
40. Collman, J. & Cooper, T. Breaching the security of the Kaiser Permanente internet patient portal: the organizational foundations of information security. *J. Am. Med. Inform. Assoc.* **14**, 239–243 (2007).
41. International HapMap Project. CEPH reconstituted form. [online]. < http://www.hapmap.org/downloads/elsj/CEPH_Reconstituted_Form.pdf >
42. Willison, D. J., Keshavjee, K., Nair, K., Goldsmith, C. & Holbrook, A. M. Patients' consent preferences for research uses of information in electronic medical records: interview and survey data. *BMJ* **326**, 373–377 (2003).
43. Baker, R., Shiels, C., Stevenson, K., Fraser, R. & Stone, M. What proportion of patients refuse consent to data collection from their records for research purposes? *Br. J. Gen. Pract.* **50**, 655–656 (2000).
44. Robling, M. R., Hood, K., Houston, H., Fay, J. & Evans, H. M. Public attitudes towards the use of primary care patient record data in medical research without consent: a qualitative study. *J. Med. Ethics* **30**, 104–109 (2004).
45. Freimer, N. B. & Sabatti, C. Variants in common disease. *Nature* **445**, 828–830 (2007).
46. Christensen, K. & Murray, J. C. What genome-wide association studies can do for medicine. *N. Engl. J. Med.* **356**, 1094–1097 (2007).
47. Hood, L., Heath, J. R., Phelps, M. E. & Lin, B. Systems biology and new technologies enable predictive and preventative medicine. *Science* **306**, 640–643 (2004).
48. Xiong, M., Feghali-Bostwick C. A., Arnett, F. C. & Zhou, X. A systems biology approach to genetic studies of complex diseases. *FEBS Lett.* **579**, 5325–5332 (2005).
49. The American Society of Human Genetics. ASHG Response to NIH on Genome-Wide Association Studies. *ASHG Policy Statement Archives* [online]. < http://www.ashg.org/pages/statement_nov3006.shtml > (2006).
50. Greely H. T. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. *Annu. Rev. Genomics Hum. Genet.* **8**, 343–364 (2007).
51. Faden, R. R. & Beauchamp, T. L. *A History and Theory of Informed Consent*. 235–241 (Oxford University Press, New York, 1986).
52. O'Neill O. *Autonomy and Trust in Bioethics*. (Cambridge University Press, Cambridge, UK, 2002).
53. Church, G., Gottlieb, D., Mitra, R. & Sherley, J. Molecular and Genomic Imaging Center. Specialized Center of Excellence in Genomic Science (CEGS) P50 Proposal *Harvard Molecular Technology Group & Lipper Center for Computational Genetics* [online]. < http://arep.med.harvard.edu/P50_03/Church03.doc > (2003).
54. Butler, D. Ghost Buster. *Nature* **445**, 811 (2007).
55. Kohane, I. S. & Altman, R. B. Health-information altruists — a potentially critical resource. *N. Engl. J. Med.* **353**, 2074–2077 (2005).
56. The GAIN Collaborative Research Group. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nature Genet.* **39**, 1045–1051 (2007).
57. Bovenberg, J. A. *Property Rights in Blood, Genes and Data – Naturally Yours? (Nijhoff Law Specials)* (Brill Academic, Leiden, 2006).
58. The subcommittee on privacy and confidentiality. Transcript of the January 28, 1998 Roundtable Discussion: Identifiability of Data. *National Committee on Vital and Health Statistics* [online]. < <http://ncvhs.hhs.gov/980128tr.htm> > (1998).
59. Motiuk, A. Anonymous sperm donor traced on internet. *New Scientist News Service* [online]. < http://www.newscientist.com/article_ns?id=mg18825244_200 > (2005).
60. The Personal Genome Project, Are guarantees of genome anonymity realistic? *Personal Genome Project* [online]. <http://arep.med.harvard.edu/PGP/Anon.htm>
61. Markoff, J. Researchers Find Way to Steal Encrypted Data. *The New York Times* [online]. <http://www.nytimes.com/2008/02/22/technology/22chip.html?ref=science> (2008).

Acknowledgements

The authors wish to thank three anonymous reviewers for their comments. J.L. thanks M. Cornel and T. Pieters of VU university medical center, Amsterdam, for discussions. R.C. gratefully acknowledges the support of the Economic and Social Research Council (ESRC). The work was part of the programme of the ESRC Centre for Economic and Social Aspects of Genomics.

Competing interests statement

The authors declare competing financial interests: see web version for details.

FURTHER INFORMATION

Jeantine Lunshof's homepage: http://www.bio.vu.nl/microb/personnel/jeantine_lunshof/index.html

American Society of Human Genetics:

http://www.ashg.org/pages/statement_nov3006.shtml

Encyclopedia of DNA Elements (ENCODE) project:

<http://www.genome.gov/10005107>

Genetic Information Nondiscrimination Act:

<http://www.genome.gov/24519851>

International HapMap Project: <http://www.hapmap.org>

Research Program on Genes, Environment and Health (RPGEH) at Kaiser Permanente:

<http://www.dor.kaiser.org/studies/rpgeh/index.html>

Mission statement of the Personal Genome Project:

<http://www.personalgenomes.org/mission.html>

NCBI Database of Genotype and Phenotype:

<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap>

The Personal Genome Project (PGP):

<http://www.personalgenomes.org>

UK Biobank: <http://www.ukbiobank.ac.uk>

ALL LINKS ARE ACTIVE IN THE ONLINE PDF