PERSPECTIVES

SCIENCE AND SOCIETY

Data sharing in genomics — re-shaping scientific practice

Jane Kaye, Catherine Heeney, Naomi Hawkins, Jantina de Vries and Paula Boddington

Abstract | Funding bodies have recently introduced a requirement that data sharing must be a consideration of all funding applications in genomics. As with all new developments this condition has had an impact on scientific practice, particularly in the area of publishing and in the conduct of research. We discuss the challenges that must be addressed if the full benefits of data sharing, as envisaged by funders, are to be realized.

The field of genomics is regarded as a leader in the development of infrastructure, resources and policies that promote data sharing[1]. Examples include the Human Genome Project and the HapMap project — which promote the sharing of sequence data — and the more recent data sharing structures for genome-wide association (GWA) studies, such as the database of Genotypes and Phenotypes (dbGaP) and the European Genotyping Archive[2]. Rapid developments in genomics are widely promoted as being dependent on such resources, which can be accessed by many researchers for different research uses. They are regarded by many as a testimony to the success of the principle of open access. In addition, all of the large funding bodies now make data sharing a requirement of support for all projects, including all hypothesis-driven projects that primarily focus on a specific research question rather than aiming to create data for the use of others. The rationale for these policies is that science and creativity are furthered by access to openly available data, and that data created by publicly funded bodies should be freely available in the research community. Even though these policies are still in their infancy, their impact is starting to be felt on the planning, execution and oversight of genomics research, and on the way in which results are disseminated.

Through our empirical work with scientists in the field[3], we have identified some key areas of scientific practice that are being affected by these policies. In this paper we discuss these four areas: the difficulties of acknowledging individual contributions to the generation of data; the way that these data sharing structures change the responsibilities of researchers towards participants; the implications that these policies have for maintaining public trust; and the new mechanisms that have been developed for oversight of access to data. These important issues illustrate the tensions that data sharing policies create for researchers, who must fulfil the requirements of funding bodies while also protecting research participants and their own career development. Failure to understand these particular tensions and the effects of these policies on scientific practice may have a detrimental impact on global goodwill and trust in genomics research, and on the development of sustainable data sharing practices. Consideration of these issues is timely, as the effects of data sharing policies are starting to be visible and understood, but are also being re-examined, as in the recent case in which genotypic data were withdrawn from internet access by the National Institutes of Health (NIH) and the Wellcome Trust[4–6].

Changes in scientific practice

The data sharing policies of funders build on and accelerate changes that have been occurring over the last two or three decades in the way that biomedical science is carried out and scientific data are generated and analysed. In genomics, change has been driven primarily by the need for fundamental sequence information, comparative populations and large numbers of samples, and by the falling costs and increasing capacity of sequencing and computing technologies. Research practice has become increasingly interdisciplinary[7], with the rapid formation of flexible and dynamic research collaborations around the world[8]. For example, the use of new methodologies in GWA studies requires: large numbers of clinically well-characterized samples to be collected from patients; laboratory staff and researchers to manage the genotyping pipeline; bioinformaticians, statisticians and other data analysts to interpret the data; and leadership from principal investigators. In combination, these factors have had a significant effect on the way that research projects are planned, organized and managed, and have encouraged the development of open access policies (BOX 1).

Hypothesis-led projects. In the case of new hypothesis-led projects, researchers are required to provide, in their funding proposal, a plan for how data and results will be shared. The specific aim of data sharing policies is to ensure maximum availability of data. Arguments can be made for excluding access to the data by some researchers on the basis of the sensitivity of the data, or the potential to identify or stigmatize individuals or groups. Although newly funded projects can be planned and developed in accordance with data sharing policies, greater challenges arise, as is indeed happening, when such policies are applied retrospectively to completed projects or to ongoing longitudinal projects.

Large-scale data generation projects. Advances in sequencing and computing technologies have also allowed the scientific community to embark on a new type of scientific effort — large-scale data generation projects. Such projects generate data and create management infrastructures, or platforms, which can support simultaneous access to a data set by multiple researchers. Secondary users of the data are far removed from the researchers

who carried out the collection of the samples and data, as well as from the research participants. In such projects, research participants are informed that the analysis of their sequence will be freely available on the Internet. These projects have enormous benefits for the entire scientific community, as they have accelerated the creation of new knowledge and provided a blueprint for data sharing (BOX 2).

The changing landscape of data sharing. The data sharing policies of funders may crystallize and encourage existing trends in scientific practice. In the past, data has primarily been shared with known colleagues and has been based on mutual respect, trust and a common interest. The conditions of access would be negotiated on an individual basis and would vary according to particular circumstances. Funders now require that data sharing be considered in every newly funded research project, unless there are justifiable reasons why this should not be so. With these policies, the question for many researchers has become how to share data, whereas previously it was whether data should be shared at all. This creates a number of challenges for several areas of scientific practice. We begin by discussing how best to provide rewards and incentives for the researchers who have been involved in data generation.

Acknowledging individual contributions

In the past, a data set would have been used primarily by the researchers who had created it, and would provide the basis for many publications. There would have been a direct relationship between the creation of the data and control over usage and the publication of results. However, with data sharing policies, the fact that particular researchers have created a data set no longer gives them an enduring priority or control over its use and resulting publications. The challenge then is how to reward and acknowledge the production of a data set.

Proper recognition for authors and contributors. The traditional form of acknowledgement is through a publication, which is also a key way of ensuring career advancement. Many journals require that data production should be acknowledged, but how this is done is largely left up to individuals, who follow the norms that exist in their particular discipline. One solution has been to publish articles with large numbers of authors[9], as this recognizes the involvement of many researchers and data producers in large collaborations. A difficulty arises, however, when the number of authors becomes excessive, as authorship is more a reflection of contribution to a project rather than to a publication. The practice adopted by some journals is to describe the contributions of individual authors, although this policy is difficult to extend to large numbers of authors. An alternative means might be to make a distinction between a 'contributor', who has provided the data set, and an 'author', who has worked on the analysis or result.

Means of recognition other than traditional authorship have also been proposed[10]. In one approach, the data set would be specifically recognized in the publication according to an established system. This would acknowledge use of the data set and indirectly reward the contributions of those who have been instrumental in establishing the resource, without needing to cite each person who contributed to the generation of the data set. Recognition in a publication is essential, but data generation needs to be established as an activity that is worthy of recognition in its own right, as it relies on specialist skills. Therefore, it is important that the efforts of data generators are appreciated by the scientific community, and that the establishment of a resource for other researchers is considered as a valuable output by institutions. In addition, there must be indices that can also be included in national assessment schemes, such as the Research Assessment Exercise in the UK, which ranks institutions according to their research excellence.

Promoting data sharing. Although publications and formal recognition are important, incentives to share data also need to be built into the research process. One solution developed by the Genetic Association Information Network (GAIN) (BOX 2), is to give the producers of the data a 6 month publishing lead on their competitors, even though the data are available to all *bona fide* researchers during this time. The researchers that generate the data are given the opportunity by funders to develop a data set using new GWA technology. However, this incentive can also place enormous pressure on the research team, who are generating data as well as attempting to analyse and publish results in a short time. Constantly working against rapidly impending deadlines is not a productive climate in the long term, and an extension of such publishing lead times should be considered. Such incentives require careful thought, as they are having a substantial effect on the way that science is being conducted, both in terms of the quality of teamwork and in the speed of data generation.

Novel ways of acknowledging contributions to the generation of data are required that are fair and transparent, lest researchers obstruct data sharing. Genomic data are

Box 1 | **Data sharing policies**

Open access to data is believed to accelerate advances in science, by making data freely available to all while ensuring the expedient use of existing resources that have been funded by the public purse. The first international document to embody this perspective and lay out the principles for open access in the field of genomics was the Bermuda Principles, agreed at the First International Strategy Meeting on Human Genome Sequencing in 1996, which was followed by the Fort Lauderdale Agreement in 2003. Together, these documents set out the key principles that now dominate thinking and practice regarding open access to genome sequence data in North America and the United Kingdom.

The key idea being promoted in the Bermuda Principles is that the pre-publication genome sequence "should be freely available and in the public domain in order to encourage research and development and to maximise its benefit to society." The Fort Lauderdale Agreement took this further by setting out a plan of tripartite responsibility for sequence producers, users and funders for the establishment of community resources to achieve rapid and open data release. This agreement stated that "community resource data sets benefit the users enormously, giving them the opportunity to analyse the data without the need to generate it first. The data sets are, in general, much larger, richer and of higher quality than individual laboratories could normally generate." Such data sets have been presented as the "drivers of progress in biomedical research" and therefore they should be "made immediately available for free and unrestricted use by the scientific community to engage in the full range of opportunities for creative science."

The open access principles underlying these developments have since been applied by national funding bodies beyond projects that generate sequence data to other areas of biomedical research. Examples of such policies are those of the National Institute of Health (2003), Genome Canada (2005) and the UK Medical Research Council (2006). All of these organizations now make data sharing a requirement of funding in genomics. These policies have created a climate in which data sharing has become the default, and applicants must demonstrate why their data should be exempt from the requirement that data should be deposited for use by other scientists.

only useful for subsequent analyses if they are accompanied by good metadata that describe, for instance, sample collection procedures, clinical definitions of the cases, and demographic data. Therefore, scientists can retain some measure of control over access[11], for example, by claiming that part of the data set is not ready to be shared. This would make it difficult for other researchers to carry out meaningful analyses[12]. This is contrary to the principles of data sharing and difficult to guard against. However, it would be inappropriate and cumbersome to develop punitive oversight mechanisms to ensure this does not happen. Instead, 'carrots' rather than 'sticks' need to be used to encourage those that create metadata to share with others further downstream in the scientific process.

Such incentives need to replicate the climate of trust and reciprocity that accompanies traditional and more informal data sharing. No one wants to be part of a system in which they feel that someone else can take advantage of their unsung contributions. One way forward is to have an open debate within the scientific community about how and why data sharing, both formal and informal, works or does not work. This debate is necessary to articulate the norms required in specific situations, and to determine a fair and equitable way to share data but also acknowledge individual contributions. This is not a matter of more regulation and guidelines, but of developing norms that become an intrinsic part of a new scientific culture, in which people can trust each other because the rules and obligations are known at the outset.

Box 2 | **Data generating projects and access criteria**

Open access policies

Several data generating projects provide free access to data online. For example, the Human Genome Project (1990–2003) aimed to sequence the 3 billion base pairs in the human genome and to identify all 20,000–25,000 genes. The HapMap project (2002–2005) identified chromosome regions with sets of strongly associated SNPs, the haplotypes in those regions and the SNPs that tag them. The 1000 Genomes Project (which began in 2007) will develop a map of biomedically relevant DNA variations at unprecedented resolution.

Each of these projects has relied on the cooperation of funders and researchers from many disciplines, and has drawn on considerable resources, expertise and time. As none of these projects provides any link to phenotypic information, access to the data is freely available through the Internet, regardless of the intended use of the data or the identity of the user.

Restricted access policies

By contrast, projects that generate, combine and archive different kinds of data, such as the database of Genotypes and Phenotypes (dbGaP), the Genetic Association Information Network (GAIN) and the Wellcome Trust Case Control Consortium (WTCCC), have developed data release policies to control access. Some data are placed on the Internet, but researchers must establish their credentials before they are allowed access to information that could potentially identify research participants.

dbGaP is a repository of four types of data: study documentation; phenotypic data; genetic data (including study subjects' individual genotypes); and statistical results, including some association and linkage analyses. dbGaP provides two levels of access — open and controlled — to allow broad release of non-sensitive data while providing oversight and investigator accountability for data sets involving personal health information. The benefit of dbGAP is that it provides a controlled archiving system for research data.

GAIN (2006–2008) completed an ambitious programme to genotype existing research studies in six major common diseases, and to combine the results with clinical data to create a substantial new research resource. The resulting data are being deposited in a database in the National Library of Medicine at the NIH, funded by GAIN, for the broad use of the research community. Originators of the initial studies received additional grants to carry out their own analyses. Access is controlled by an NIH data access committee.

The WTCCC (2007–present) is a collaboration of 24 geneticists based in the United Kingdom that is analysing thousands of DNA samples from patients to identify common genetic variations for different diseases. Aggregated data are placed on the Internet, but access to the more detailed genotypic and phenotypic data is obtained only through the principal investigator, who can also decide on further collaboration.

The primary goal of all these initiatives is to make data as widely available as possible to further scientific progress. However, decisions about access are centralized and are no longer controlled by the research team who collected the data; instead researchers must conform to specific deposition and access requirements, which affect the way in which research is conducted.

Responsibilities towards study participants

The original context in which the samples and data are collected is associated with expectations and relationships that are understood both by researchers and participants[13]. Researchers may feel a strong sense of responsibility for 'their' samples and feel a legal and moral responsibility for research participants that often extends beyond the original terms of consent. This responsibility may not be felt by secondary researchers who have no connection with the research participants, and see themselves as only dealing with data. Although secondary researchers have an obligation to use data in a scientifically sound, ethical and lawful manner, these obligations are not the same as the researchers enrolling patients in a study. Informed consent forms, which try to be succinct, may not embody all of the expectations that are associated with enrolment in a study and an ongoing clinical relationship, and may leave room for differing interpretations of the scope of consent.

In data sharing policies, researchers are given the opportunity to justify why raw data should not be shared. Given their knowledge about the types of uses that may be made of data, based on the original consent, researchers are in a good position not only to decide on appropriate uses, but also to protect against possible misuse. In particular, when samples are collected and analysed as an extension to ongoing epidemiological work, cohort studies, or disease-specific work in which the relationship develops in a clinical setting, the obligation to share genomic data may be perceived as an imposition on the relationships that have been built up between researchers and participants.

The challenge for funders is to ensure that this sense of stewardship is respected, by ensuring that new systems for sharing data acknowledge these perceived responsibilities. There is a danger that data sharing policies may be experienced as being punitive, or that those who feel uncertain about sharing may be characterized as obstructive and short-sighted. However, reluctance to share may have sound justifications; such concerns cannot be ignored, as they can have practical as well as ethical implications in a project in which the trust and support of participants is vital. In addition, researchers who are perceived as uncooperative could be excluded from key areas of activity, such as developing strategic policies and being involved in peer-review.

Funding bodies must be prepared to consider the claims of those for whom data sharing, because of the nature of their research and situation, may create difficulties. At the moment, requests for exemption from data sharing are judged by funders, but it may be better for this assessment to be made by bodies that are independent from considerations about future funding for the applicants.

Box 3 | **Consent**

Models such as broad consent have been proposed as a solution to some of the ethical challenges of data sharing. In broad consent, an individual gives consent to widely specified research, which allows for many future uses of tissue and data rather than just the one (or more) use(s) specified by known researchers. Once individuals give consent, they are not re-contacted concerning new uses. In projects that contain uncertainty about the scope of the consent, authorization for the use of coded samples and data may be given by a research ethics committee.

However, there is concern about whether this practical solution to the issue of informed consent is compliant with data protection principles[27], which require that the individual should know how, and by whom, their data are being processed.

There is also concern that broad consent undermines one of the fundamental principles of medical research, that of individual autonomy and the right of individuals to decide the nature of their involvement in medical research[28]. However, there are differing conceptions of autonomy. In some views, individual autonomy requires decisions to be based on full information. According to others, full information is not required for autonomous consent as long as individuals understand the broad nature of what is proposed and understand that they do not have all the details of what is involved. The latter situation demands a greater level of trust in the individuals and institutions concerned.

Maintaining public trust

The mechanisms that have traditionally been used to protect research participants are informed consent and the anonymization of data sets. However, the sharing of data from genomic studies tests the effectiveness of these standard mechanisms of privacy protection.

Anonymization of data. The digital revolution, which has allowed many types of data to be shared both with and without consent, is rapidly changing the landscape of privacy protection. Procedures for controlling disclosure, such as coding each study subject or aggregating the information, can be employed to protect the identity of data subjects. However, these policies may lessen the scientific utility of the data, as fine detail and nuances can be lost in the effort to protect privacy[14]. Furthermore, as DNA is a unique identifier, it is impossible to completely anonymize a sample, and small numbers of SNPs can be used to identify individuals[15]. The recent decisions by the Wellcome Trust and the NIH to remove SNP data from publicly accessible databases, following a paper by Homer *et al.*[16], illustrate the problems of protecting participants' privacy interests when using GWA methodology. Homer and his colleagues established by a statistical analysis that an individual could be identified in aggregate data, as genome-wide scans provide such a wide range of unique data points.

Informed consent. The process of obtaining informed consent is one way for research participants to have some control over how their information is used. However, this procedure is problematic when it is applied in a data sharing context.

First, it is difficult to achieve the level of understanding that is required for truly informed consent[17–19], especially for data sharing in genomics: participants have a variable understanding of whether their sequence data will be shared, and with whom[20]. Second, it is difficult to provide information about all the potential users of shared data, without a constantly updated system to inform participants. Many long-term studies, such as the Avon Longitudinal Study (ALSPAC) have approached this problem through websites for participants that enhance an understanding of the science. Greater patient involvement in the decision making of biobanks has also been proposed to compensate for this deficit[21] (BOX 3).

Data sharing challenges existing mechanisms for privacy protection. Once data have been released into the public domain, participants and researchers have little or no control over usage of the data, or the possibility that they may be linked to other data sets. Research participants can exercise only consent or withdrawal; it is difficult for participants to control how their genomic data will be shared — typically they are required to consent either to all data being shared between researchers or no sharing at all. In addition, there are real doubts about whether an individual's request for withdrawal can be meaningful, owing to the complexity of retracting data that has been used in different data sets. In this new context of global data sharing, better methods of informing participants about the use of their personal information for different research purposes need to be developed.

Oversight of access

Data sharing raises new dilemmas for the oversight of research and for the bodies that are entrusted to ensure that research is well governed. Traditionally, approval for research is obtained from a research ethics committee by a particular individual or research group. This committee holds the principal applicant responsible for monitoring the use of samples and data; however, when samples are transported across national borders, and when data are analysed by people who bear no relation to the original research project or participants, it is almost impossible to continue to hold the original collector responsible in the same way. Therefore, it is difficult for research ethics committees to exert their original mandate to ensure the ethical conduct of research.

Box 4 | **Global data sharing**

Organizations such as Public Population Project Genomics Consortium (P3G), the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) and the Organization for Economic Cooperation and Development (OECD) have started the legal analysis that is required to develop mechanisms that promote global data sharing while ensuring that research is carried out ethically and according to accepted standards. Ideally, the new framework would relieve researchers from having to seek approval from multiple data access committees, but this direction is still being debated. One possibility is the development of a system in which approval for access is given by one international body for a number of similar projects, rather than having independent access committees for each project. This could develop uniformity in decision making, and create a clear and transparent set of criteria for deciding questions of access for all researchers. The disadvantage is that it removes decision making from the local level to a body that is removed from the context in which the data set has been established.

One of the problems of such a proposition is that, although international agreements can help to set broad standards, all countries have their own systems of law. This means the flow of data and samples through a number of countries will be subject to many different legal regimes, and to different sets of guidelines and standards. The concept of an international body to oversee data sharing is good in theory, but in reality it would probably add another layer of bureaucracy for researchers, as they would be forced to comply with the international layer of approval as well as comply with national regulations.

It has recently been suggested that ethics committees could have an important function in monitoring the particular ethical aspects of GWA studies. Although such responsibilities have mainly focused on the consent process[22] and on approving the re-use of samples[23], some have also considered the potential contribution that ethics committees could make to regulating data access[24,25]. However, ethics committees are already facing increasing challenges in reviewing complex research proposals[26], and it is therefore not clear how already overburdened committees could take on the role of monitoring and approving data access — a task that requires significant insight into the techniques used to produce and analyse data in genomics.

Instead of relying on research ethics committees, some data sharing initiatives have established specialist advisory bodies, or 'data access committees', to determine who should have access to data and on what grounds. The criteria used for making these decisions are still in their infancy and can vary among projects. However, many of the publicly available criteria for determining access seem to involve establishing whether a scientist is a '*bona fide* researcher', rather than considering whether access will have an effect on research participants. The criteria to qualify as a *bona fide* researcher are still being developed, but they could include type of research or could simply be based on verifying an individual's credentials, such as institutional affiliation. In the case of dbGaP at the NIH, users are granted a data use certificate, which allows access to several data sets. Although this generic authorization may be efficient, there is the possibility that the privacy interests of research participants may be overlooked when a number of data sets are linked together (BOX 4).

Conclusions

Data sharing policies have been in place in the United States for the past 6 years and in the United Kingdom for approximately 3 years, and their full effect on scientific practice is just starting to be understood. The challenge for individual researchers is to simultaneously fulfil the requirements of funding bodies, honour their obligations to study participants and protect their own interests and careers. The challenge for funders is to ensure that public trust is maintained, and that data sharing policies improve the transfer of research results and knowledge. To ensure that this framework is sustainable, good working relationships are required between funders, collaborating groups of scientists and the many thousands of recruits that will be needed in the future.

Establishing and maintaining global public goodwill and trust is an ongoing task to ensure the future of sound, and hence ethical, scientific research. Meeting such challenges is necessary to ensure that data sharing practices and policies continue to produce the harvest of tangible benefits currently enjoyed by some of the scientific community. However, further research and thought is required to address some of the challenges to scientific practice that data sharing creates.

Jane Kaye, Catherine Heeney, Naomi Hawkins, Jantina de Vries and Paula Boddington are at The Ethox Centre, Department of Public Health and Primary Care, University of Oxford, Old Road Campus, Headington, Oxford OX3 7LF, UK.

Correspondence to J.K.
e-mail: jane.kaye@ethox.ox.ac.uk

doi:10.1038/nrg2573
Published online 24 March 2009

1. The Digital Archiving Consultancy (DAC), The Bioinformatics Research Centre, University of Glasgow (BRC), The National e-Science Centre (NeSC). *Large-scale Data Sharing in the Life Sciences: Data Standards, Incentives, Barriers and Funding Models (The "Joint Data Standards Study").* [online], < http://www.nesc.ac.uk/technical_papers/UKeS-2006-02.pdf > (2005).
2. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9**, 356–369 (2008).
3. Gibbons, S. M. C., Kaye, J., Smart, A., Heeney, C. & Parker, M. Governing genetic databases: challenges facing research regulation and practice. *J. Law Soc.* **34**, 163–189 (2007).
4. Couzin, J. Genetic privacy. Whole-genome data not anonymous, challenging assumptions. *Science* **321**, 1278 (2008).
5. [No authors listed]. DNA databases shut after identities compromised. *Nature* **455**, 13 (2008).
6. Kaiser, J. House weighs proposal to block mandatory 'open access'. *Science* **321**, 1621 (2008).
7. Ziman, J. *Real Science: What It Is, and What It Means* (Cambridge Univ. Press, Cambridge, 2000).
8. Gibbons, M., Limoges, C., Nowotny, H., Schwartzmann, S. S. P. & Trow, M. *The New Production of Knowledge. The Dynamics of Science and Research in Contemporary Societies* (Sage Publications, London, 1994).
9. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
10. Cambon-Thomsen, A. Assessing the impact of biobanks. *Nature Genet.* **34**, 25–26 (2003).
11. Campbell, E. G. *et al.* Data withholding in academic genetics: evidence from a national survey. *JAMA* **287**, 473–480 (2002).
12. Foster, M. W. & Sharp, R. R. Share and share alike: deciding how to distribute the scientific and social benefits of genomic data. *Nature Rev. Genet.* **8**, 633–639 (2007).
13. Nissenbaum, H. Protecting privacy in an information age: the problem of privacy in public. *Law Philos.* **17**, 559 –596 (1998).
14. Karp D. R. *et al.* Ethical and practical issues associated with aggregating databases. *PLoS Med.* **5**, e190 (2008).
15. McGuire, A. L. & Gibbs, R. A. Genetics: no longer de-identified. *Science* **312**, 370–371 (2006).
16. Homer, N. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4**, e1000167 (2008).
17. Lunshof, J. E., Chadwick, R., Vorhaus, D. B. & Church, G. M. From genetic privacy to open consent. *Nature Rev. Genet.* **9**, 406–411 (2008).
18. McGuire, A. L., Caulfield, T. & Cho, M. K. Research ethics and the challenge of whole-genome sequencing. *Nature Rev. Genet.* **9**, 152–156 (2008).
19. Caulfield, T. *et al.*, Research ethics recommendations for whole-genome research: consensus statement. *PLoS Biol.* **6**, e73 (2008).
20. McGuire A. L., Hamilton J. A., Lunstroth R. J. D., McCulloch L. B. & Goldman, A. DNA data sharing: research participants' perspectives. *Genet. Med.* **10**, 46–53 (2008).
21. Winickoff, D. E. Partnership in U.K. Biobank: a third way for genomic property? *J. Law Med. Ethics* **35**, 440–456 (2007).
22. Mascalzoni, D., Hicks, A., Pramstaller, P. & Wjst, M. Informed consent in the genomics era. *PLoS Med.* **5**, e192 (2008).
23. Muula, A. S. & Mfutso-Bengo, J. M. Responsibilities and obligations of using human research specimens transported across national boundaries. *J. Med. Ethics* **33**, 35–38 (2007).
24. Upshur, R., Lavery, J. & Tindana, P. Taking tissue seriously means taking communities seriously. *BMC Med. Ethics* **8**, 11 (2007).
25. Langat, S. K. Reuse of samples: ethical issues encountered by two institutional ethics review committees in Kenya. *Bioethics* **19**, 537–549 (2005).
26. Hoedemaekers, R., Gordijn, B., Hekster, Y. & Van Agt, F. The complexities of ethical evaluation of genomics research. *HEC Forum* **18**, 18–36 (2006).
27. Zika, E. *et.al.* Sample and data use and protection in biobanking in Europe: legal issues. *Pharmocogenomics* **9**, 773–781 (2008).
28. Caulfield T. Biobanks and blanket consent: the proper place of the public perception and public good rationales. *King's Law J.* **18**, 209–226 (2007).

Acknowledgements
We would like to thank the anonymous reviewers for their insightful comments and suggestions. J.K., C.H., N.H. and J. de V. are funded by the Wellcome Trust, under grant codes WT 081407/Z/06/Z, WT 076070/Z/04/Z, WT 077869/Z/05/Z and WT 083326/Z/07/Z. P.B. is funded by EU FP6, Procardis Project number 037273.

FURTHER INFORMATION

The Ethox Centre: http://www.ethox.org.uk
1000 Genomes Project: http://www.1000genomes.org/page.php
Avon Longitudinal Study (ALSPAC): http://www.bristol.ac.uk/alspac/participants
Bermuda Principles: http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml#1
Biobanking and Biomolecular Resources Research Infrastructure (BBMRI): http://www.bbmri.eu/bbmri
Database of Genotypes and Phenotypes (dbGaP): http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html
European Genotyping Archive: http://www.ebi.ac.uk/ega/page.php
Fort Lauderdale Agreement: http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf
Genetic Association Information Network (GAIN): http://fnih.org/index.php?option=com_content&task=view&id=338&Itemid=454
Genome Canada Data Release & Resource Sharing Policy: http://www.genomecanada.ca/medias/PDF/EN/DataReleaseandResourceSharingPolicy.pdf
HapMap project: http://www.hapmap.org/abouthapmap.html
Human Genome Project: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
National Institute of Health Data Sharing Policy and Implementation Guidance: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm
Organization for Economic Cooperation and Development (OECD): http://www.oecd.org
OECD Guidelines for Human Biobanks and Genetics Research Databases: http://www.oecd.org/document/12/0,3343,en_2649_34537_40302092_1_1_1_1,00.html
Public Population Project Genomics Consortium (P3G): http://www.p3gconsortium.org
UK Medical Research Council policy on data sharing and preservation: http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Datasharinginitiative/Policy/index.htm
Wellcome Trust Case Control Consortium (WTCCC): http://www.wtccc.org.uk

ALL LINKS ARE ACTIVE IN THE ONLINE PDF