Navigating the perfect [data] storm

M.J. Murtagh, G.A. Thorisson, S.E. Wallace, J. Kaye, I. Demir, I. Fortier, J.R. Harris, D. Cox, M. Deschênes, P. Laflamme, V. Ferretti, N.A. Sheehan, T.J. Hudson, A. Cambon Thomsen, R.P. Stolk, B.M. Knoppers, A.J. Brookes[¥] and P.R. Burton[¥], on behalf of the P^{3}G Consortium, GEN2PHEN and BioSHARE-EU

[¥] joint senior authors

Correspondence: Madeleine Murtagh, Department of Health Sciences, University of Leicester, Room 209, Adrian Building, University Road, Leicester LE1 7RH, United Kingdom

E-mail: mm399@leicester.ac.uk Telephone: +44 (0)116 252 2926 Telefax: +44 (0)116 252 3748

ABSTRACT

Bioscience has recently undergone a series of knowledge-based and technological revolutions. A critical consequence has been increasing recognition of the need to invest in infrastructure. Good access to data (and samples) from multiple studies is axiomatic to the value of this infrastructure. Access must be streamlined, secure, and based upon transparent and 'fair' decision making. It must be clear who has created and who has used which data. Ethico-legal policies and guidelines, which reflect dominant local cultural and societal norms, must take account of the increasingly global nature of bioscience research. A robust data infrastructure must also be attentive to the translational aims and social impact of its knowledge generation. In order to maintain the trust of its constituency – the general public as well as professional, political, commercial stakeholders – it must develop mechanisms to take account of all of these perspectives. These considerations form the basis of an emerging *data economy*. Building on extant achievements and pursuing the ideas outlined here could revolutionise the way we use and manage large-scale data. They have critical implications for biomedical and public health research communities and will be of central relevance for healthcare managers and policy makers, governments and industry. However, if the major challenges are to be met we must continue to invest, both nationally and internationally, in developing the cooperative infrastructures that provide a complementary foil to competitive resourcing mechanisms that drive hypothesis-driven science.

INTRODUCTION

Scientific advance involves the asking and answering of questions within constraints of contemporaneous knowledge and technology. Until recently, most definitive 'answers' in health science reflected factors with relatively large effects (*e.g.* the health impact of smoking cigarettes). But, the study of the etiological architecture of common chronic diseases demands that we explore much weaker effects (*1,2*) including interactions (*3,4*). This poses obvious challenges for statistical power (*5*). Moreover, weak relationships are easily created or concealed by confounding or reverse-causality (*6*). Provision of an effective platform for tomorrow's biomedical science therefore demands high quality data on an unprecedented scale. Furthermore, many research questions necessitate co-analysis of multiple studies, placing a premium on data harmonization (*7*) and stream-lined access. Though the shape of this emergent data economy (or, more accurately, economi*es*) is as yet unclear, its evolution is rapidly gathering momentum.

We are facing a 'perfect [data] storm' on four main fronts which need resolution to enable the development of an effective platform for biomedical science. First, there is a need to create political, legal and ethical frameworks for data governance that incorporate privacy issues and protect research participants' personal information, whilst also being attentive to the ethical dimensions of scientific enterprise, such as intellectual property rights and recognition of the investment of scientists (*8,9*). Second, there is a need to establish effective mechanisms for recognising the substantive contributions of *everybody* in building, maintaining and operating data infrastructures (*8,10*), not just the research leaders that obtain funding (*11*). Third, there is a need to optimise the exploitation of an increasing deluge of large, complex data sets (*12-14*) and to identify the social dimensions of optimising data curation (*15*) and data sharing (*16,17*). Fourth, the management and use of data and the generation of knowledge needs to be taken forward with social impact and translational aims in mind, particularly by engaging the insights of all relevant stakeholders (*16*).

Although the storm is already upon us (*9*), it was forecast by international organisations including P^{3}G (*18,19*), BBMRI (*20,21*), ISBER (*22,23*), PHOEBE (*7,24*) and GEN2PHEN (*25*) working in the field of population genomics. Solutions have been developed by these initiatives for some of the most pressing issues (*7*): study cataloguing (*18,21,26*); data harmonization (*7,27*); and, ethico-legal, social and political issues underpinning data management and access (*28,29*). This paper highlights these solutions and additional endeavours that could dramatically change how we manage future data, not only in bioscience or -omics research but across domains using large-scale potentially shareable data (cf. FlaReNet (*30*) and Gigascience (*31*)).

GOVERNING DATA ETHICALLY

Protecting participants

The maintenance of public and scientific trust in the systems of scientific governance is fundamental to successful data sharing. Technological advances must work within the existing political, legal and cultural environment to have legitimacy and be socially accepted. One of the challenges is to build governance structures that allow the free movement of data to encourage scientific advancement while at the same time ensuring that individual participant's data are protected from harm. The ethical, epistemological, social and practical barriers to data-sharing within the research community need to be studied, practical solutions need to be researched, and changes implemented. Innovative solutions need to be developed to address the confidentiality and anonymisation challenges that arise from the rich, complex and potentially identifiable data that can be amassed through biomedical infrastructure. Changing and diverse societal attitudes towards privacy as well as new ways of engaging research participants through social media and IT solutions need to be incorporated into the development of new governance frameworks.

While research is increasingly global with large international, multidisciplinary collaborations and studies that span national borders, our current regulatory mechanisms for research are nationally based. Working to create frameworks at an international level (c.f. P^3G generic consent materials for population biobanks) for adaptation within local settings can help to address this issue. Likewise, frameworks developed at the local level can inform international policies. Considerable work has been done already to link up national endeavours with international platforms and to co-ordinate efforts through organisations such as P^3G and ISBER, to prospectively harmonize these efforts and meet the future challenges of e-governance. However, the accelerating pace of genome science will demand an internationally coherent approach if we are to have any chance to address future challenges. And, as we argue below, this necessitates the active consideration of the viewpoints of a range of stakeholders; whether scientific, professional, public or participant (*16*). Development of such a global vision for ethical, legal and the social implications (ELSI) of genomics is underway (*32*) and must underpin protection both of the participants, whose data are the basis of scientific knowledge, and of the scientists and others who produce that knowledge.

Identifying scientific contributions

Transparent identification of data and their origins is central to the acknowledgment of *all* contributions in the ideal data management infrastructure. This requires all actants (*33*) (material or human entities) to be unambiguously, computationally and securely identifiable. In practice, this would mean assigning digital identifiers (IDs) and sometimes version numbers to everything, not least: biobanks/cohorts; the institutions that host resources; research participants; datasets and databases; scientists that generate the 'dataverse'; individuals/organisations that use the stored and shared information and journals that publish their findings. Currently, only some of these carry IDs, but optimal data-sharing and usage will not be achieved until such IDs become ubiquitous, properly designed, and widely recognised and used.

There is potential to leverage online digital IDs to establish a globally distributed and seamlessly automated system for facilitating data access – bringing benefits of speed, transparency, and equity (*8*). The scheme (*34*) (Box 1), would greatly improve current processes for granting access to potentially sensitive datasets. Several small scale projects under the auspices of GEN2PHEN and BioSHaRE-EU are currently piloting controlled access to summary-level, aggregate datasets aiming to roll out this approach for use with more sensitive data. Such a system would, for example, have circumvented the data release controversies that followed Homer et al. (*35*). It would also ameliorate the current hindrance of scientific progress by delays and complications involved in gaining access to this class of data – a situation at odds with the obligation to maximize the knowledge generated by publicly-funded research (*8*).

ANALYSING DATA THAT CANNOT BE ACCESSED

Social and ethico-legal imperatives driving expectations of security, privacy and transparency have already engendered important changes in how we use, share and analyse data. For example, when data are physically very large, Kahn (*12*) argues that streamlined analysis may benefit from moving "computation to the data, rather than the data to the computation". But, this idea can be taken an important step further; enabling the secure *joint* analysis of data from *several* studies, even when some of those studies are unable to share raw data. This is crucial because conventional approaches to joint analysis cannot optimise the efficiency and flexibility of the statistical analysis whilst simultaneously ensuring that all relevant ethico-legal and governance stipulations are met in full (Box 2). DataSHIELD provides a novel solution to this challenge (*36*).

Under DataSHIELD (Figure 1 and Box 2) *full joint analysis* is achieved via simultaneous parallelized analysis of the individual-level data at each study. The approach is iterative and – at each iteration – the separate parallel analyses are linked by exchanging summary statistics with the analysis centre. These statistics carry no sensitive information and are non-identifying; in these regards they are equivalent to the study-level results that are shared freely under SLMA (study level meta-analysis – see Box 2). Furthermore, although the analysis is *mathematically equivalent to ILMA* (individual level meta-analysis – see Box 2), the participant-

Box 1. Unique IDs for researchers and bio-resources.

Researcher IDs

- ORCID
  - The Open Researcher and Contributor ID initiative (*53,54*) is constructing a global registry of unique, permanent and institutionally verifiable IDs for authors of scholarly publications.
  - ORCID will enable reliable disambiguation of one author from another, plus new knowledge capabilities discovery mediated via searching across these unambiguous IDs.
- ORCID (extended)
  - Extension of the ORCID concept into the online world of databases and data sharing could meet the goal of appropriately recognising (and ultimately rewarding) the intellectual and other inputs of researchers to construction, maintenance and use of all aspects of the global data and information infrastructure.
  - Unambiguous identification of individual researchers and science contributors could provide the foundation of a rapid, IT mediated access mechanism (*34*) for data of low or moderate disclosure risk, that cannot be posted freely on the web but would ideally be available over the web with light touch oversight control – a data access approach termed "speed pass". Given broad-based acceptance by the research community, a willingness of institutions to be responsible for their own *bona fide* scientists and recognition that proscribed misuse of data or samples might lead to loss of access rights, such a system would, for example, provide an ideal response to the limited risk of identification posed by the methods described by Homer et al. *(35)*.

Bioresource IDs

- BRIF
  - The Bioresource Research Impact Factor (*55,56*) has been proposed as an indicator of the use of all bio-resources (biobanks, cohorts, reference collections and databases).
  - Each bio-resource should have its own internationally unique and persistent recognised ID, as a necessary step to automatically trace its use.
  - Standardisation of citation using this ID is required .
  - It would facilitate tracking of the contribution of individuals to bio-resources, and of bio-resources to bioscience and to the bioscience infrastructure as a whole.
  - An international working group (18) is currently addressing the various dimensions of such a tool.

Box 2. Two conventional approaches to jointly analysing (meta-analysing) multiple studies.

- **Two conventional approaches to joint analysis**
  - (i) Study level meta-analysis (SLMA): investigators at each study analyse their own data; they return results to a central analysis centre (AC); results are meta-analysed at the AC.
  - (ii) Individual level meta-analysis (ILMA): individual level data (de-identified) are physically transferred from each study to the AC; data from all studies are analysed together.
- **Choice of approach from the perspective of the science and statistical analysis**
  - SLMA works if analysis can be completely pre-planned and if it is straightforward to specify and obtain the study level results that are required.
  - ILMA is greatly to be preferred if any exploratory analysis is required. Every unplanned analysis under SLMA causes serious delay as each group of study investigators must re-analyse their own data and return the new results to the AC.
- **Choice of approach from the ethico-legal perspective**
  - Individual level data cannot physically be transferred if governance materials (consent forms, information leaflets, conditions applied by an ethics committee) prohibit it.
  - Even when the transfer of individual level data is permitted, it is likely to require a lengthy access process involving scientific oversight and ethics committees. The pace of progress in contemporary bioscience is such that research groups fear losing out to competitors.
- **How should we move forward?**
  - An approach is needed that allows timely meta-analysis of individual-level data but avoids the need for data to be physically transferred, or even visible, outside of the original study in which they were collected. DataSHIELD (*36*) is such an approach.

level data never leave their study of origin and remain invisible to the analysing statistician. Given appropriate ethico-legal consent, therefore, use of DataSHIELD might arguably be permissible even if the collaborating studies are prohibited from physically sharing data (*29*).

MAINTAINING INTEGRITY

Scientific, technical and ethico-legal mechanisms can only facilitate data sharing where there can be an assurance of their integrity and of the outcome of the data sharing. In turn, it is the ultimate production of

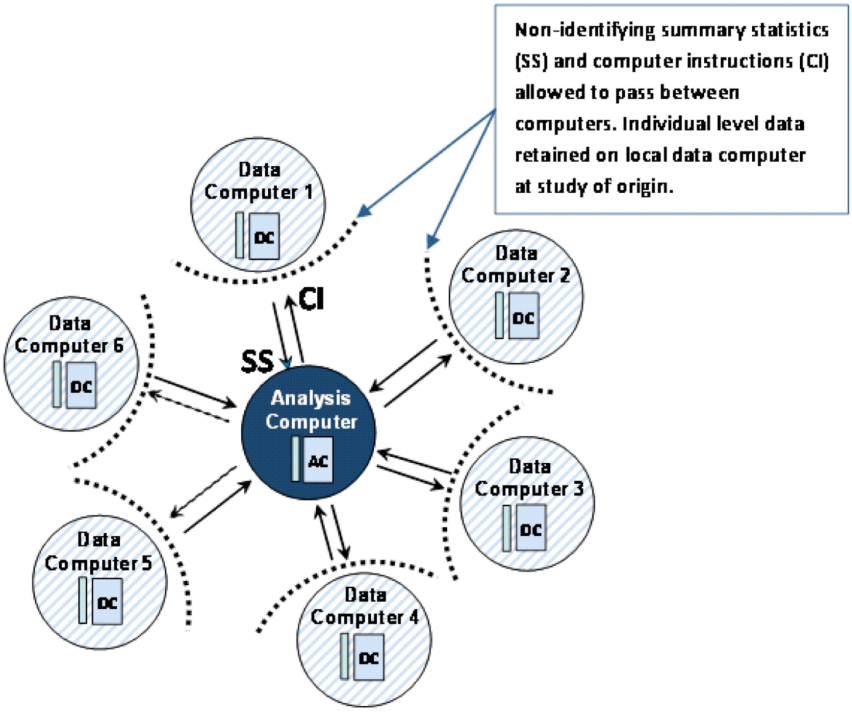

Figure 1. Schematic overview of IT architecture for DataSHIELD as applied to six studies. Analysis Computer (dark shading) runs R(*51*). Data Computers (light diagonal stripes) run OPAL(*52*) and R. Each Data Computer linked with Analysis Computer over internet via firewalls.

publically and politically acceptable translational outcomes of scientific knowledge that provides 'definitive' evidence of scientific integrity. In the context of data sharing, trust is fundamental to maintaining integrity. Trust functions at a number of levels: between participants and scientists in the collection of data; between scientists (from all active disciplines) in the production of knowledge; between stakeholders – scientific, political, commercial and public – in the application of outputs of the knowledge generated. The acceptability of science and its products is effectively driven by trust (*37,38*). Development of trust requires active engagement and such engagement must be fit for purpose and tailored to its members. Engaging the public, participants, scientists and other stakeholders serves at least two purposes: maintaining public and participant trust in the science and scientific process; contributing public and stakeholder views and perspectives to the development of that science. Arguably, therefore, a key function for engagement is to ensure attention to the translational aims and social impact of scientific knowledge. Engagement is, therefore, a tool for translation – i.e. $T1_{tm}$ – in translational science terminology (*16*).

Translation of scientific knowledge into societal impact (health and health service improvement) requires development of tools and mechanisms for the strategic engagement of stakeholders. *Hybrid forums*, that is, discussions incorporating transdisciplinary, multisector representation *(39)*, based on existing international collaborations, have the potential to transcend disciplinary and science/professional boundaries and barriers, thereby fostering communication and trust. International transdisciplinary groups of natural, social and humanities scientists are already established (*e.g.* P^3G Observatory (*19*), BBMRI (*20*)). Funded appropriately, these groups could form the basis of extended forums, to include political, policy, commercial and professional stakeholders, for integrated strategic discussion about developments in the science, its translation and social impact. Further, issues of trust – central for the effective production of scientific knowledge – must be acknowledged. Increased specialization, collaborations and teamwork within science, scientific activity today necessitates that scientists assess and trust the integrity of their colleagues – whether this activity is data collection, data processing, experimentation, interpretation of results, or peer review. Trustworthiness of other members of scientific community is a central foundation of scientific knowledge generation; what sociologists and historians of science describe as the *epistemic* role of trust (*40-43*). While trust can enhance and aid cooperation, interaction and sharing, lack of trust can hamper not only the production of knowledge but also its effective exchange and sharing. Thus, enhancing effective data sharing in biomedical sciences will need to take into consideration the social and practical processes which impact upon trust between scientists. This requires social science research to identify, describe and reflect upon those barriers and their impact on knowledge production.

Engaging the public and research participants argu-

ably requires different methods (*44,45*). Involving members of the public in conventional governance and organisational meetings is the most common mode of engagement. But this engagement can be tokenistic, may include only the 'usual suspects' or those with known views and may not therefore result in the incorporation of the valuable insights that may be derived from pubic perspectives (*44,46*). Public meetings and consultations specifically addressing issues in genomics may predominantly attract those motivated by extreme views (*47*). Resistance to the disempowering characteristics of conventional engagement in these settings can lead to aetiolated outcomes (*48*). Ironically, these approaches risk undermining rather than maintaining public trust. A strategy for genuinely engaging the public must be multifaceted: it must comprise individuals as well as communities; be purposive as well as using evolutionary mechanisms for engagement; it must take advantage of new Web 2.0 social media; and must target existing forums with broad appeal. However, engaging motivated individuals and groups is not the same as gaining a genuine insight into public perceptions. If we really want to understand public views of specific issues in genomics and biobanking, for example privacy, we need to undertake well conducted, appropriately designed research to do so (cf. *49,50*). In other words, understanding public views and perceptions requires robust, theoretically-informed and adequately resourced social science research. Only in this way can we properly inform the development of socially appropriate and acceptable scientific knowledge generation.

"WHAT'S PAST IS PROLOGUE"

(The Tempest, 2.1)

Bioscience has recently undergone a series of knowledge-based and technological revolutions. A critical consequence has been increasing recognition of the need to invest in infrastructure. Good access to data (and samples) from multiple studies is axiomatic to the value of this infrastructure. Access must be streamlined, secure, and based upon transparent and "fair" decision making (*8*). It must be clear who has created and who has used which data (*10*). Ethico-legal policies and guidelines, which already reflect local cultural and societal norms, must take account of the increasingly global nature of bioscience research (*32*). A robust data infrastructure must also be attentive to the translational aims and social impact of its knowledge generation (*16*). In order to maintain the trust of its constituency – the general public as well as professional, political, commercial stakeholders – it must develop mechanisms to take account of all of these perspectives.

But this is no *tabula rasa*. Despite its obvious benefits and regardless of the approach used, shared data analysis must conform to long-standing principles: for example, analysis is simply not valid unless the studies to be combined are *harmonized (7)*; likewise, harmonized data sets will be useless if data from one study cannot be shared beyond national borders because data governance requirements and policies do not allow it. Building on extant achievements and pursuing the ideas outlined here could revolutionise the way we use and manage large-scale data. They have critical implications for biomedical and public health research communities and will be of central relevance for healthcare managers and policy makers, governments and industry. However, if the major challenges are to be met we must continue to invest, both nationally and internationally, in developing the cooperative infrastructures that provide a complementary foil to competitive resourcing mechanisms that drive hypothesis-driven science.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the contribution of the Public Population Project in Genomics (P^3G), GEN2PHEN and BioSHARE-EU. In addition, the authors acknowledge the following sources of support: MJM, SW, ID, AJB, PRB are members of the methodological and infrastructure research programme, Data to Knowledge for Practice, at the University of Leicester which is funded jointly under the BioSHaRE-EU project (European Commission, FP7, #261433), Wellcome Trust Supplementary Grant #086160/Z/08/A, and joint MRC/Wellcome Trust Project Grant #G1001799/#WT095219MA. The DataSHIELD and Opal software development at the Ontario Institute for Cancer Research is funded under the BioSHaRE-EU project (European Commission, FP7, #261433). JRH gratefully acknowledges support for this work through funds from the European Union's Seventh Framework Programme (FP7/2007-2013), ENGAGE Consortium, grant agreement HEALTH-F4-2007-201413; BioSHaRE-EU, grant agreement HEALTH-F4-2010-261433; and through funds from Biobank Norway – a national infrastructure for biobanks and biobank related activity in Norway – funded by the Norwegian Research Council (NFR 197443/F50). BMK is supported by a Canada Research Chair in Law and Medicine. TJH and VF are recipients of Investigator Awards from the Ontario Institute for Cancer Research, through generous support from the government of Ontario. BMK and TJH receive funding support for the Public Population Project in Genomics (P^3G) from Genome Canada, Genome Quebec, and the European Commission (ENGAGE, FP7-Health-201413). JK is supported by Wellcome Trust 096599/2/11/Z. NAS is supported by Leverhulme Trust Research Fellowship RF/9/RFG/2009/0062. AJB is supported by UK Medical Research Council (COPDmap), EU FP7 integrated projects (GEN2PHEN (grant # 200754), and BioShaRE (grant #261433)).

REFERENCES

1. 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 (2007).
2. T.A. Manolio, L.D. Brooks, F.S. Collins, A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* **118**, 1590 (2008).
3. T.A. Manolio, J.E. Bailey-Wilson, F.S. Collins, Genes, environment and the value of prospective cohort studies. *Nature Rev Genet* **7**, 812 (2006).
4. F. Kauffmann, A. Cambon-Thomsen, Tracing biological collections: between books and clinical trials. *JAMA* **299**, 2316 (2008).
5. P.R. Burton *et al.*, Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol* **38**, 263 (2009).
6. G. Taubes, Epidemiology faces its limits. *Science* **269**, 164 (1995).
7. I. Fortier *et al.*, Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol* **39**, 1383 (2010).
8. M. Walport, P. Brest, Sharing research data to improve public health. *Lancet* **377**, 537 (2011).
9. G. King, Ensuring the data-rich future of the social sciences. *Science* **331**, 719 (2011).
10. E. Pisani, C. AbouZahr, Sharing health data: good intentions are not enough. *Bull WHO* **88**, 462 (2010).
11. S. Staff, Challenges and opportunities. *Science* **331**, 692 (2011).
12. S.D. Kahn, On the future of genomic data. *Science* **331**, 728 (2011).
13. H. Akil, M.E. Martone, D.C. Van Essen, Challenges and opportunities in mining neuroscience data. *Science* **331**, 708 (2011).
14. T. Rowe, L.R. Frank, The disappearing third dimension. *Science* **331**, 712 (2011).
15. D. Howe *et al.*, Big data: The future of biocuration. *Nature* **455**, 47 (2008).
16. M.J. Murtagh, I. Demir, J.R. Harris, P.R. Burton, Realizing the promise of population biobanks: a new model for translation. *Hum Genet* **130**, 333 (2011).
17. G.A. Thorisson, Accreditation and attribution in data sharing. *Nature Biotechnol* **27**, 984 (2009).
18. B.M. Knoppers, I. Fortier, D. Legault, P. Burton, The Public Population Project in Genomics (P^3G): a proof of concept? *Eur J Hum Genet* **16**, 664 (2008).
19. P^3G Observatory (http://www.p3gobservatory.org, 2009).
20. BBMRI (http://www.bbmri.eu/, 2011).
21. H.E. Wichmann *et al.*, Comprehensive catalog of European biobanks. *Nature Biotechnol* **29**, 795 (2011).
22. ISBER, Best practices for repositories I: Collection, storage, and retrieval of human biological materials for research. *Cell Preserv Technol* **3**, 5 (2005).
23. ISBER (http://www.isber.org/, 2011).
24. PHOEBE (www.phoebe-eu.org/, 2011).
25. Gen2Phen (http://www.gen2phen.org/, 2011).
26. M.D. Mailman *et al.*, The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* **39**, 1181 (2007).
27. P.J. Stover, W.R. Harlan, J.A. Hammond, T. Hendershot, C.M. Hamilton, PhenX: a toolkit for interdisciplinary genetics research. *Curr Opin Lipidol* **21**, 136 (2010).
28. K. Hoyer, The ethics of research biobanking: a critical review of the literature. *Biotechnol Genet Eng Rev* **25**, 429 (2008).
29. S. Wallace, S. Lazor, B.M. Knoppers, Consent and population genomics: The creation of generic tools. *IRB* **31**, 15 (2009).
30. FLaReNet (http://www.flarenet.eu/, 2011).
31. Gigascience (http://www.gigasciencejournal.com/, 2011).
32. J. Kaye, E.M. Meslin, B.M. Knoppers, E.T. Juengst, Global ELSI – A Research Strategy for Genomics (http://www.p3g.org/secretariat/events/GlobalELSI%20Research210412.pdf; http://www.publichealth.ox.ac.uk/helex/news/developing-a-global-vision-for-the-future-of-elsi-research, 2011).
33. B. Latour, *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, Mass.: Harvard University Press, 1987.
34. P^3G Consortium, Public access to genome-wide data: five views on balancing research with privacy and protection. *PLoS Genet* **5**, e1000665 (2009).
35. N. Homer *et al.*, Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* **4**, e1000167 (2008).
36. M. Wolfson *et al.*, DataSHIELD: resolving a conflict in contemporary bioscience – performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol* **39**, 1372 (2010).
37. M. Dixon-Woods, R.E. Ashcroft, Regulation and the social licence for medical research. *Med Health Care Philos* **11**, 381 (2008).

38. M. Dixon-Woods, D. Wilson, C. Jackson, D. Cavers, K. Pritchard-Jones, Human tissue and 'the public': the case of childhood cancer tumour banking. *BioSocieties* **3**, 57 (2008).
39. M. Callon, C. Méadel, V. Rabeharisoa, The economy of qualities. *Economy and Society* **31**, 194 (2002).
40. H.M. Collins, Tacit knowledge, trust and the Q of sapphire. *Soc Stud Sci* **31**, 71 (2001).
41. T.M. Porter, *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press, 1996.
42. S. Shapin, *A social history of truth: Civility and science in seventeenth-century England*. Chicago: University of Chicago Press, 1994.
43. I. Demir, Lost in translation? Try second language learning: Understanding movements of ideas and practices across time and space. *J Hist Sociol* **24**, 9 (2011).
44. S. Weldon, Public engagement in genetics: a review of current practice in the UK (NOWGEN Report). Lancaster: Lancaster University (2004).
45. M. Learmonth, G.P. Martin, P. Warwick, Ordinary and effective: the Catch 22 in managing the public voice in health care? *Health Expect* **12**, 106 (2009).
46. S. Peacock *et al.*, Using economics to set pragmatic and ethical priorities. *Circulation* **108**, 697 (2003).
47. P. Burton, Power to the people? How to judge public participation. *Local Economy* **19**, 193 (2004).
48. M.J. Murtagh, Enagement as disempowerment (Forthcoming).
49. H. Gottweis (http://private-gen.eu/, 2011).
50. G. Gaskell, H. Gottweis, Biobanks need publicity. *Nature* **471**, 159 (2011).
51. R Development Core Team, *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing (2008).
52. OBiBa (http://www.obiba.org/, 2011), vol. 2009.
53. M. Fenner, C.G. Gómez, G.A. Thorisson, Key Issue: Collective Action for the Open Researcher & Contributor ID (ORCID) *Serials: The Journal for the Serials Community* **24**, 277 (2011).
54. ORCHID (http://www.orcid.org, 2011).
55. A. Cambon-Thomsen, Assessing the impact of biobanks. *Nature Genet* **34**, 25 (2003).
56. A. Cambon-Thomsen *et al.*, The role of a Bioresource Research Impact Factor as an incentive to share human bioresources. *Nature Genet* **43**, 503 (2011).