Genome Medicine

CORRESPONDENCE

Developing and implementing an institute-wide data sharing policy

Stephanie OM Dyke and Tim JP Hubbard*

Abstract

The Wellcome Trust Sanger Institute has a strong reputation for prepublication data sharing as a result of its policy of rapid release of genome sequence data and particularly through its contribution to the Human Genome Project. The practicalities of broad data sharing remain largely uncharted, especially to cover the wide range of data types currently produced by genomic studies and to adequately address ethical issues. This paper describes the processes and challenges involved in implementing a data sharing policy on an institute-wide scale. This includes questions of governance, practical aspects of applying principles to diverse experimental contexts, building enabling systems and infrastructure, incentives and collaborative issues.

Introduction

The Wellcome Trust Sanger Institute (WTSI) played an important role in the international public effort to sequence the human genome, the Human Genome Project (HGP), which has become a symbol of the benefits of policies on early release of scientific data. The HGP data release policy, known as the 'Bermuda Agreement', was agreed to in 1996 by a group of genomic scientists and funders that included leaders from WTSI and the Wellcome Trust, and built on successful practices that had been in operation in other fields of genetics (for example, the *Caenorhabditis elegans* Genome Project [1-3]). Other WTSI sequencing projects, whose structure easily fits the specifics of the HGP data release policy, followed suit and adopted similar practices that rapidly became WTSI policy [4]. Large-scale international collaborations, such as the SNP Consortium [5], Mouse Genome Sequencing Consortium [6] and International HapMap Project [7], also decided to follow HGP practices and to share data publicly as a resource for the research community before academic publications describing analyses of the data sets had been prepared (referred to as prepublication data sharing).

Following the success of the first phase of the HGP [8] and of these other projects, the principles of rapid data release were reaffirmed and endorsed more widely at a meeting of genomics funders, scientists, public archives and publishers in Fort Lauderdale in 2003 [9]. Meanwhile, the Organisation for Economic Co-operation and Development (OECD) Committee on Scientific and Technology Policy had established a working group on issues of access to research information [10,11], which led to a Declaration on access to research data from public funding [12], and later to a set of OECD guidelines based on commonly agreed principles [13]. These initiatives, and those of other fora, firmly established data sharing as a priority in the minds of individuals involved, and in particular led to the development of funders' policies in the UK and USA [14-17].

However, by 2003 genomic science had diversified with a range of different data types being collected across multiple species. Funders were beginning to look at standards for large-scale data in other fields of the life sciences [18]. As WTSI shifted focus from a few large sequencing projects to multiple endeavors, coordination on data sharing for studies that involved different funders, different technologies and diverse institutions became increasingly complex. Efforts to maintain the principles associated with HGP data release therefore led to a range of project-specific adaptations. This approach worked well for large-scale studies that had sufficient resources to manage data sharing plans, such as The Encyclopedia of DNA Elements (ENCODE; 2003 and 2008 [19,20]), Wellcome Trust Case Control Consortium (WTCCC; 2005 [21]), Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER; 2006 [22]), 1000 Genomes Project (2008 [23]), International Cancer Genome Consortium (ICGC; 2008 [24]) and MalariaGen (2008 [25]), but led to disparities in adherence to data sharing for smaller projects.

*Correspondence: th@sanger.ac.uk
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Furthermore, projects were starting to use human data sets that engendered additional ethical considerations. As it became possible to study genomic data for large numbers of individuals, the genomics community, with its evolving data sharing standards, began to interact more with the human genetics community, whose practices placed greater emphasis on data confidentiality. It became accepted that a reasonable way to ensure the benefits of data sharing, while managing the risks, was to share data with controls to limit access to approved users for approved purposes. In 2006, a purpose-built 'managed access' database, the database of Genotypes and Phenotypes (dbGaP), was established in the USA for storing and sharing genotypes and associated phenotypes that could not be published through existing public archives [26]. In 2007, a similar repository was set up at the European Bioinformatics Institute (EBI): the European Genome-phenome Archive (EGA) [27]. WTSI has continued to actively participate in relevant policy discussions with the Wellcome Trust and other funders, such as the Toronto International Data Release Workshop in 2009, which led to the development of the Toronto Statement [28].

In summary, at the same time as these complexities evolved, it became more widely accepted that increased data sharing was important. It has become recognized that data sharing enables research, accelerates translation, safeguards good research conduct, and helps inform policy and regulation, thereby fostering a public climate in which research can flourish. Being committed to these benefits spurred the Institute to develop and implement an institute-wide data sharing policy.

Developing and implementing the policy

A review of data sharing policy at WTSI, including a consultation to identify issues of concern, was undertaken. This allowed an institute-wide data sharing policy to be drafted that covers the diverse work being carried out. A working group that included faculty members representing every area of WTSI science was set up to steer this effort. The process of review and policy revision took a year and the drafting of policy followed a standard course that has been described previously [29].

The policy that resulted from this process addresses ethical issues and differences in experimental contexts and data types [30]. It includes a commitment to rapid sharing of data sets of use to the research community (which include primary and processed data sets, research articles and software code), and encompasses elements to address the following: (1) protection of research participants; (2) promotion of respect for rights for data generators of acknowledgement and first publication; (3) provisions to facilitate translation into health benefits; (4) fair access procedures; (5) transparency (with respect to availability of data as well as of access procedures); (6) adoption of recognized data and interoperability standards, including submission to designated public repositories.

For many aspects of data sharing policy, best practice for implementation remained to be established. While carrying out the review of data sharing policy, the Institute began to devote resources to support the implementation of the Wellcome Trust policy on open and unrestricted access to research articles (in brief: papers describing research carried out at or in collaboration with WTSI must be made publicly available through UK PubMed Central (UKPMC) as soon as possible and in any event within 6 months of the journal publisher's official date of final publication [31]). This effort focused on the development of 'how-to-comply' guidelines, including information for collaborators [32] and instituting records of submissions and compliance tracking, with support from research administrators and library staff. Based on this experience, it was agreed that successful policy implementation would depend on working out detailed requirements (guidance), devoting efforts and resources to alleviate disincentives (facilitation), instituting monitoring processes (oversight), and leadership. These are discussed in detail below in the following sections: Guidance, Facilitation and Oversight.

Guidance

A major challenge was to work out what the principles outlined in the text of the policy meant in practice for individual projects. Decisions were guided by the need to ensure that anticipated benefits from making data available would outweigh the costs associated with long-term archiving and the effort involved in preparing data for submission. Timelines for submission were determined by evaluating the length of time required to allow adequate quality control to ensure value over time. For example, reference genome sequence data are valuable with minimal quality control. The value of the draft human genome sequence data shared within 24 h of sequencing is testament to this approach. On the other hand, certain cellular assays captured through sequencing (for example, ChIP-seq) may have little value if the experiment failed and this may not be realized until initial analysis has been carried out.

The appropriate resolution of raw data submitted was also considered in this way. Summary data sets can be much smaller than the raw data sets they derive from, and in many cases satisfy the needs of other users. On the other hand, storing raw data is more important if samples are rare or where methods to summarize data are still in development. These considerations affect the decisions about what data to archive, and they may change over time. For example, for submission of next-generation sequence data, the guidance has changed over the last

year from sequence read format (SRF) to binary sequence alignment/map format (BAM) [33]. Over this period it has become accepted in the community that the value of the extra information stored in SRF format related to sequence quality has diminished as methods have become more standardized. In addition, the mapping information contained within the BAM format makes the files more easily reused without further processing (see Discussion). Since the cost of generating sequence data continues to fall rapidly, there are already discussions about further reducing the amount of stored information [34].

Relatively specific guidelines for different data/study types were therefore developed that were nevertheless generic enough to apply to very different experiments. For example, functional analysis assays were grouped as one category even though they involve different data types and even different technologies. This was because of similar requirements for greater quality control (as described above) and similar lower anticipated value of raw data sets to others. However, within this category, transcriptomics data sets were felt to be of broader use, because of the likelihood that they contained novel expressed sequence, and were therefore set to be shared earlier. Target timelines for the submission of primary and processed data sets of different data/study types were generally set following this kind of reasoning. Finally, suitable public repositories and data formats for submission were identified, with a view to enhancing data reuse through ease of discovery and ease of integration with other data sets.

It was also necessary to define procedures for the handling of and access to 'managed access' data sets that could not be shared without restrictions to protect confidentiality and the privacy of research participants, or to respect the terms of their consent. Managing access to data sets involves determining who may access the data and for what purpose(s) through an application process and setting out conditions of data access in a data access agreement. This therefore involved preparing a standardized data access agreement that provided sufficient protection while allowing maximal reuse and outlining data security parameters for the use of 'managed access' data sets. Associated guidance has also been developed for access to research articles (as described above) and for software releases.

It was important that an initial version of the data sharing guidelines be circulated at the time of the policy first being published. This facilitated the development of the guidelines document through further discussion/consultation with scientists across the Institute. One of the initial drivers for this work was to ensure consistency in policy application. Developing a suitable framework was an iterative process, incorporating feedback and experience from individual projects. Regular and honest communication of the policy development process that was being undertaken, along with strong leadership, allowed for support to be maintained throughout the year that it took to establish a working version of the guidelines, which remain under constant review. Ultimately, this led to consensus guidelines that were developed from the bottom up, and this influenced subsequent adoption across the Institute. As soon as they were reasonably fit for purpose, a public version of the data sharing guidelines was published on WTSI website [35].

Facilitation

In terms of disincentives, the issues identified during the consultation process fell into two main categories: concerns about the difficulty of rapidly sharing data effectively because it is time-consuming, technically difficult and involves taking responsibility for access decisions; and concerns about credit (mainly with respect to scientific competition and protection of rights of first publication and of intellectual property).

Data sharing, especially on a large scale, is still difficult and time consuming. WTSI decided that it would not serve as a data repository wherever suitable public repositories had been established for particular data types or scientific fields. It was recognized that data sets available from central repositories are easier to discover and integrate with other data sets, thereby enhancing data reuse. In addition, storing and making data available has significant cost implications for an institute and creates a long-term obligation that may become disconnected from research interests. WTSI therefore committed core resources to assist researchers with many of the time-consuming/technical steps involved in submitting data to the designated repositories, such as metadata collation. Processes were automated wherever feasible and project managers and research administrators trained so that they could help develop plans and facilitate submission.

Integrating data pipelines and tools across WTSI research programs (including planning the development of shared data resources wherever needed) has allowed the Institute to enhance the efficiency and cost-effectiveness of important steps in the data sharing process. For the data types that WTSI researchers produce on a very large scale, namely next-generation sequencing data sets, a substantial investment was made to develop automatic submission pipelines to the three major databases that would be their destination: the European Nucleotide Archive (ENA) [36], the EGA [27] and Array Express (AA; [37]) (Figure 1). Cooperation and coordination with EBI, especially over metadata standards, has been essential to achieve this, in particular for newer data types such as RNA-seq (where standards are still being developed [38]). Supporting systems such as these is

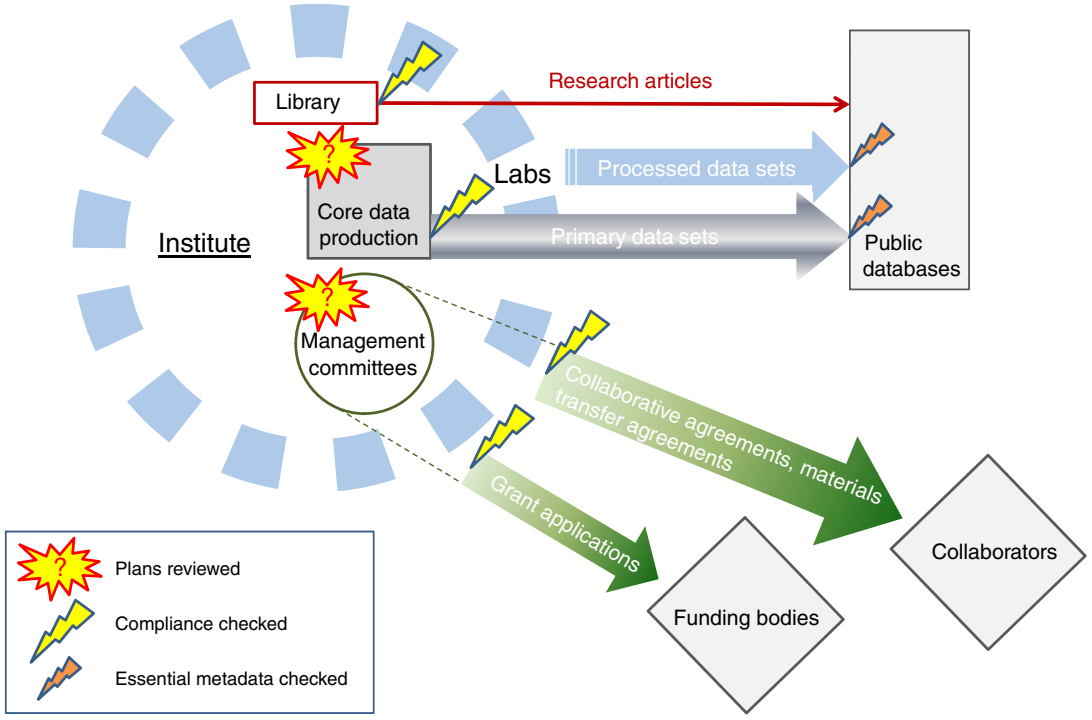

Figure 1. Monitoring data sharing plans. The processes involved in monitoring both plans and practice in institute data sharing. Checkpoints that occur within management committees and within software systems that handle data submissions are highlighted. Primary sequencing data sets are submitted through an automatic pipeline.

costly, but justifiable, for an institute producing data on a large scale and it has dramatically improved the process of data sharing, the quality and consistency of submissions, and overall compliance.

A key aspect to successful data sharing is that researchers need to be relatively confident that users of the data will respect conditions of data access, especially rights of first publication upon which the success of their careers can depend. Publication moratoria aim to ensure that researchers sharing data before they have published research articles describing their analysis are still able to do so. They prohibit publications by others that would deprive data generators of credit, while ideally still allowing publication of non-competing analysis. Publication moratoria are effectively a codification of the principles outlined originally in the report of the Fort Lauderdale meeting [9]. ENCODE and the ICGC are two large-scale research consortia whose data sharing policies include publication moratoria [20,24]. Standard data access 'conditions of use' statements were therefore developed, both incorporating principles adopted elsewhere (for example, publication moratoria that are both defined in scope and time-limited) and through the formulation of new concepts such as the 'data display' agreement, developed for the DECIPHER project [22]. The 'data display' agreement allows DECIPHER data to be integrated into third party web displays through a requirement that the data be presented in such a way that conditions of use are respected, and this includes notifying users of the obligations on them [39]. Users wishing to analyze the full DECIPHER 'managed access' data set would have to be approved and agree to the data access agreement for the project.

WTSI is also trying to promote data sharing etiquette through more prominent communication of expectations on its website and with data submissions. Website developments such as central listings of data available have also enhanced the discovery of data resources. For example, the data resource pages were reorganized to provide a structured catalog of genome data sets linked to accessions in repository databases [40]. This led to an observed marked increase in web accesses to this area.

Oversight

In order to oversee policy developments and institute systems for monitoring data sharing plans and practices, the data sharing working group was established as a governance body. It was decided that monitoring should be proactive, strike the right balance between control-based and trust-based approaches, and build on existing mechanisms of oversight wherever possible. Committee members adopted a flexible approach for projects that

had been established prior to the policy update and until the guidelines were sufficiently refined.

Data sharing has been fully integrated into WTSI planning processes. The policy update coincided with WTSI quinquennial strategic review and this allowed the scientific research programs to develop data sharing plans (requested as part of the review process) that were consistent with the policy. In addition, standard internal forms, used for approval of external grant applications and registration of internal projects, had data sharing questions added to them. These allow data sharing plans to be checked and defined early on in the research process (Figure 1). WTSI's network of management committees raised awareness of the policy through review of data sharing plans submitted with project applications.

Another important aspect of implementation has been to ensure that any legal and other collaborative agreements are compatible with the policy by reviewing them with this in mind (for example, material transfer agreements, data transfer/access agreements, research collaboration agreements). The introduction of standardized clauses into these agreements has reduced the workload associated with this review. Having these template documents in place, alongside the data sharing guidelines, has helped WTSI researchers communicate default WTSI expectations to collaborators. It has also been important to ensure that data sharing plans are consistent with the expectations of research participants and to better communicate our data sharing expectations, and in some cases risks, to individuals involved in studies and to the ethics bodies reviewing research plans.

Several tools that were extended to facilitate submission of data sets to the public archives have the additional benefit of allowing practices to be overseen. For example, the project management software package Sequencescape that was developed in-house for the production of large-scale data sets captures instructions used by the automatic submission pipelines described previously (Figure 1). When setting up projects using Sequencescape, users select data sharing options corresponding to their data sharing plans. The information recorded allows WTSI to produce and check reports on data sharing practices.

Discussion

Looking back on our experiences, we believe that in order to be effective, data sharing policy implementation needs to be carried out in a systematic and comprehensive way, such as described here. Given the constant pressures on researchers, it is easy for data sharing to be seen as a burden, and neglected. Much of this work has been to reduce this burden by both clarifying exactly how to go about data sharing and facilitating it. While implementation takes time, our experience is that these processes have already significantly improved the ability of WTSI to share data rapidly. Much of this progress has been achieved in the context of work within high-profile multi-institutional projects that have established standards, and through ownership of the policy by faculty members, scientific managers and others, especially those closely involved in the review. The Wellcome Trust has also always provided invaluable leadership through its data sharing policy initiatives. Furthermore, regular discussions with the Wellcome Trust have allowed practical difficulties encountered at an institutional level to be addressed, an example being the allocation of additional resources to handle decisions on access requests for 'managed access' data sets. A few of the current outstanding issues are now discussed.

Cultural barriers to data sharing continue to exist, as reasons not to share can seem to outweigh the benefits and community norms have not been fully established [41,42]. It is therefore important to promote data sharing by demonstrating its benefits (see examples below) and aligning reward systems to ensure that scientists sharing data are acknowledged/cited [43,44] and that this activity is credited in research assessment exercises and grant/career reviews. The publication moratorium system, whereby scientists share data with the understanding that users will not publish analyses within a given area, has helped encourage early data submission; however, it will take time to assess its overall effectiveness. One danger of moratoria is unintentionally delaying analyses by other groups and this is one reason why time limits on moratoria are important. Institute efforts can address these challenges to some extent, as has been recommended by Piwowar *et al*. [45]; however, funders, publishers and public archives have an important role to play [45], especially in clarifying and communicating agreed etiquette and in developing responses to abuses of the system [46]. A declaration upon publication stating that users have abided by any conditions of data access, similar to the recently introduced conflict of interest statements, would help ensure these conditions are respected.

At WTSI, investigators are responsible for archiving most processed data types in appropriate repositories. The requirements of journals create a strong incentive, and several journals have recently reinforced and extended their policies on data access [47-49]. These developments are being driven in part by the growing recognition of the importance and difficulties of ensuring reproducibility in modern fields of enquiry involving large data sets and computational analysis [50,51].

It is essential that the entire scientific community of researchers and funders is satisfied of the overall benefit of data sharing to science. The potential of data reuse to advance science is not fully explored, nor are the wider benefits of data sharing [52]. However, there are examples

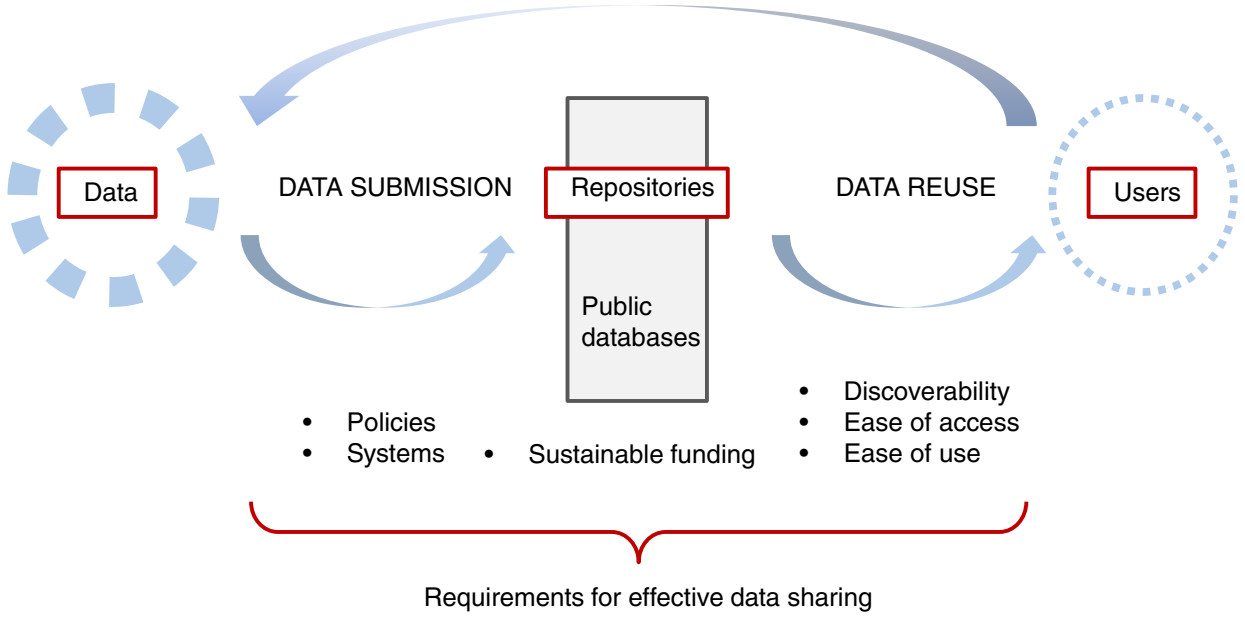

Figure 2. The data sharing ecosystem. The main requirements for effective data sharing. For data sharing to function, the processes of submission, archiving and access for reuse must all be optimized. If the barriers to any step are too high, the full benefits of data sharing will not be realized.

where benefits can be directly demonstrated. For example, the Framingham Heart Study [53] data have led to 2,223 research articles. Clinical and imaging data collected for the Alzheimer's Disease Neuroimaging Initiative [54] had by February 2011 provided the basis for 160 papers, with at least 80 more to come [55]. One study provides evidence that articles on cancer microarrays for which raw data are shared are cited 70% more frequently than those that do not [56]. It is widely recognized that breakthroughs in many areas of science depend on the integration and analysis of very large amounts of shared data. However, it is clear from the evolution of DNA sequence archive policy (described above) that the cost/benefit of data archiving needs to be kept under review with respect to the resolution that is preserved, particularly where technology is changing rapidly. There are currently insufficient metrics to allow the value of data submissions of different qualities to be assessed. Indeed it is hard to quantify the reuse of any data set with no robust mechanism for capturing the data dependencies of research articles.

Despite the developments described here, the requirements for science based on large-scale data generation, sharing and reuse are still evolving. For example, it is clear that effective data sharing is dependent on more than data submission alone (Figure 2). Repositories need to be adequately funded to support archiving the increasing volumes of data. The increasing importance of research infrastructures to support the handling and storage of large-scale data has been recognized under the roadmap process set up by the European Strategic Forum for Research Infrastructures (ESFRI) [57]. In addition, repositories must ensure that discovering and accessing archived data sets is easy enough to encourage exploration without becoming a disproportionate maintenance burden. A promising recent strategy is the adoption of submission formats for nucleotide data that contain the mapping to a reference genome (for example, the BAM format mentioned above [33,58]). Genome browsers that support these formats [59-61] can federate such data sets on-the-fly without even downloading the file from the archive. This degree of ease of use makes it practical for researchers to browse data sets speculatively.

Finally, there is currently broad interest in cross-discipline data linking, partly stimulated by government initiatives to make raw data available to encourage the development of new analysis and services to improve society [62]. In the field of medical research it has been recognized that clinical applications of genomics will become important in clinical practice, as discussed in the recent UK House of Lords report on Genomic Medicine [63]. Linking genetic data to electronic health records and government data sets will facilitate analysis that should lead to improved healthcare treatments and provision. Clearly, increased data sharing enables this, though where data sets require 'managed access', data linking is inherently more complex to ensure data security and privacy are maintained.

Conclusions

The historical mode of scientific communication, including that of data, has been through scientific collaboration and journal publication. In today's world of massive data sets and of almost unlimited computational resources, there is a huge potential to accelerate science through increased data sharing, independent of formal collaboration or publication. However, while data sharing may be in the interests of society, in the competitive world of scientific research, data sharing does not just happen. In this paper we have outlined our experiences in facilitating increased data sharing at an institutional level and the issues that still remain.

Abbreviations

BAM, binary sequence alignment/map format; DECIPHER, Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources; EGA, European Genome-phenome Archive; EBI, European Bioinformatics Institute; ENCODE, The Encyclopedia of DNA Elements; HGP, Human Genome Project; ICGC, International Cancer Genome Consortium; OECD, Organisation for Economic Co-operation and Development; SRF, sequence read format; WTSI, The Wellcome Trust Sanger Institute.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SD and TH contributed equally to the design and preparation of the manuscript.

Authors' information

SD is Policy Adviser at WTSI. TH is Head of Informatics at WTSI, and Chair of WTSI Data Sharing Committee.

Acknowledgements

The authors are grateful to members of WTSI Data Sharing Committee, Faculty and WTSI staff who have supported developments in the implementation of data sharing policy. The authors would also like to thank Wellcome Trust and European Bioinformatics Institute staff, and consortium collaborators, for their support. The preparation of this manuscript was supported by the Wellcome Trust grants 079643 and 077198.

Published: 28 September 2011

References

1. **Summary of Principles Agreed at the First International Strategy Meeting on Human Genome Sequencing: 25-28 February 1996.** Bermuda HUGO; 1996. [http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml]
2. Bentley D: **Genomic sequence information should be released immediately and freely in the public domain.** Science 1996, ***274:***533-534.
3. Waterston R, Sulston J: **The genome of *Caenorhabditis elegans*.** *Proc Natl Acad Sci U S A* 1995, **92:**10836-10840.
4. **Sanger Institute Data Release Policy (1998).** [http://web.archive.org/web/19980625053324/www.sanger.ac.uk/Projects/release-policy.shtml]
5. The International SNP Map Working Group: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature* 2001, **409:**928-933.
6. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, *et al.*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
7. International HapMap Consortium: **The International HapMap Project.** *Nature* 2003, **426:**789-796.
8. International Human Genome Sequencing Consortium: **The publication of the working draft of the human genome by the International Human Genome Sequencing Consortium: Initial Sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
9. **Sharing data from large-scale biological research projects: a system of tripartite responsibility.** Report of a meeting organized by the Wellcome Trust and held on 14-15 January 2003 at Fort Lauderdale, USA. [http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf]
10. Arzberger P, Schroeder P, Beaulieu A, Bowker G, Casey K, Laaksonen L, Moorman D, Uhlir P, Wouters P: **Science and government. An international framework to promote access to data.** *Science* 2004, **303:**1777-1778.
11. **Promoting Access to Public Research Data for Scientific, Economic, and Social Development.** OECD Follow Up Group on Issues of Access to Publicly Funded Research Data, Final Report; 2003. [http://dataaccess.ucsd.edu/Final_Report_2003.pdf]
12. OECD: **OECD Declaration on Access to Research Data from Public Funding.** Adopted on 30 January 2004 in Paris.
13. **OECD Principles and Guidelines for Access to Research Data from Public Funding.** [http://www.oecd.org/dataoecd/9/61/38500813.pdf]
14. **National Institutes of Health Data Sharing Policy.** [http://grants.nih.gov/grants/policy/data_sharing/]
15. **Medical Research Council policy on data sharing and preservation.** [http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Datasharinginitiative/Policy/index.htm]
16. **Wellcome Trust policy on data management and sharing.** [http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm]
17. **Biotechnology and Biological Sciences Research Council Data sharing policy.** [http://www.bbsrc.ac.uk/organisation/policies/position/policy/data-sharing-policy.aspx]
18. The Digital Archiving Consultancy, The Bioinformatics Research Centre (University of Glasgow) and The National e-Science Centre: **Large-scale data sharing in the life sciences: data standards, incentives, barriers and funding models (The 'Joint Data Standards Study').** [http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC002552]
19. The ENCODE Project Consortium: **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science* 2004, **306:**636-640.
20. The ENCODE Project Consortium, Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, Wold B, Crawford GE: **A user's guide to the encyclopedia of DNA elements (ENCODE).** *PLoS Biol* 2011, **9:**e1001046.
21. The Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447:**661-678.
22. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett RM, Carter NP: **DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources.** *Am J Hum Genet* 2009, **84:**524-533.
23. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061-1073.
24. International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusada J, Lane DP, Laplace F, Youyong L, Nettekoven G, Ozenberger B, Peterson J, Rao TS, Remacle J, Schafer AJ, Shibata T, Stratton MR, *et al.*: **International network of cancer genome projects.** *Nature* 2010, **464:**993-998.
25. The Malaria Genomic Epidemiology Network: **A global network for investigating the genomic epidemiology of malaria.** *Nature* 2008, **456:**732-737.
26. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST: **The NCBI dbGaP database of genotypes and phenotypes.** *Nat Genet* 2007, **39:**1181-1186.
27. **The European Genome-phenome Archive.** [http://www.ebi.ac.uk/ega/]
28. Toronto International Data Release Workshop Authors, Birney E, Hudson TJ, Green ED, Gunter C, Eddy S, Rogers J, Harris JR, Ehrlich SD, Apweiler R, Austin CP, Berglund L, Bobrow M, Bountra C, Brookes AJ, Cambon-Thomsen A, Carter NP, Chisholm RL, Contreras JL, Cooke RM, Crosby WL, Dewar K, Durbin R, Dyke

SO, Ecker JR, El Emam K, Feuk L, Gabriel SB, Gallacher J, Gelbart WM, *et al.*: **Prepublication data sharing.** *Nature* 2009, **461:**168-170.
29. Field D, Sansone SA, Collis A, Booth T, Dukes P, Gregurick SK, Kennedy K, Kolar P, Kolker E, Maxon M, Millard S, Mugabushaka AM, Perrin N, Remacle JE, Remington K, Rocca-Serra P, Taylor CF, Thorley M, Tiwari B, Wilbanks J: **'Omics data sharing.** *Science* 2009, **326:** 234-236.
30. **Wellcome Trust Sanger Institute Data Sharing Policy.** [http://www.sanger.ac.uk/datasharing/]
31. **Wellcome Trust Open Access Policy.** [http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Open-access/index.htm]
32. **Wellcome Trust Sanger Institute Publication Policy.** [http://www.sanger.ac.uk/datasharing/docs/wtsi_publication_policy.pdf]
33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **2:**2078-2079.
34. Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E: **Efficient storage of high throughput DNA sequencing data using reference-based compression.** Genome Res 2011, **21:**734-740.
35. **Wellcome Trust Sanger Institute Data Sharing Guidelines.** [http://www.sanger.ac.uk/datasharing/docs/wtsi_datasharing_guidelines.pdf]
36. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, Jang M, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Sobhany S, Ten Hoopen P, Vaughan R, Zalunin V, Cochrane G: **The European Nucleotide Archive.** *Nucleic Acids Res* 2011, **39:**D28-31.
37. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A: **ArrayExpress - a public database of microarray experiments and gene expression profiles.** *Nucleic Acids Res* 2007, **35:**D747-750.
38. **BioSharing.** [http://otter.oerc.ox.ac.uk/biosharing/]
39. **DECIPHER v5.1 data sharing policy.** [http://decipher.sanger.ac.uk/datasharing/]
40. **Wellcome Trust Sanger Institute website data resource pages.** [http://www.sanger.ac.uk/resources/downloads/]
41. Campbell EG, Clarridge BR, Gokhale M, Birenbaum L, Hilgartner S, Holtzman NA, Blumenthal D: **Data withholding in academic genetics: evidence from a national survey.** *JAMA* 2002, **287:**473-480.
42. Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P: **Data sharing in genomics - re-shaping scientific practice.** *Nat Rev Genet* 2009, **10:**331-335.
43. **Data producers deserve citation credit [editorial].** *Nat Genet* 2009, **41:**1045.
44. Cambon-Thomsen A, Thorisson GA, Mabile L, Andrieu S, Bertier G, Boeckhout M, Cambon-Thomsen A, Carpenter J, Dagher G, Dalgleish R, Deschênes M, di Donato JH, Filocamo M, Goldberg M, Hewitt R, Hofman P, Kauffmann F, Leitsalu L, Lomba I, Mabile L, Melegh B, Metspalu A, Miranda L, Napolitani F, Oestergaard MZ, Parodi B, Pasterk M, Reiche A, Rial-Sebbag E, Rivalle G: **The role of a Bioresource Research Impact Factor as an incentive to share human bioresources.** *Nat Genet* 2011, **43:**503-504.
45. Piwowar HA, Becich MJ, Bilofsky H, Crowley RS; caBIG Data Sharing and Intellectual Capital Workspace: **Towards a data sharing culture: recommendations for leadership from academic health centers.** PLoS Med 2008, **5:**e183.
46. Guttmacher AE, Nabel EG, Collins FS: **Why data-sharing policies matter.** *Proc Natl Acad Sci U S A* 2009, **106:**16894.
47. Hanson B, Sugden A, Alberts B: **Making data maximally available.** *Science* 2011, **331:**649.
48. **Standard cooperating procedures [editorial].** *Nat Genet* 2011, **43:**501.
49. Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG: **Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers.** *BMJ* 2010, **340:**181.
50. Kleppner D, Sharp PA: **Research data in the digital age.** *Science* 2009, **325:**368.
51. Stodden V: **The scientific method in practice: reproducibility in the computational sciences (February 9, 2010).** MIT Sloan School Working Paper no. 4773-10 [http://ssrn.com/abstract=1550193]
52. Boulton G, Rawlins M, Vallance P, Walport M: **Science as a public enterprise: the case for open data.** *Lancet* 2011, **377:**1633-1635.
53. **Framingham Heart Study.** [http://www.framinghamheartstudy.org/index.html]
54. **Alzheimer's Disease Neuroimaging Initiative.** [http://www.adni-info.org/]
55. Travis K: **Sharing data in biomedical and clinical research.** *Science (Career Magazine)* 2011. doi: 10.1126/science.caredit.a1100014.
56. Piwowar HA, Day RS, Fridsma DB: **Sharing detailed research data is associated with increased citation rate.** *PloS One* 2007, **2:**e308.
57. **European Roadmap for Research Infrastructures, Roadmap 2008.** [http://ec.europa.eu/research/infrastructures/pdf/esfri_report_20090123.pdf]
58. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D: **BigWig and BigBed: enabling browsing of large distributed datasets.** *Bioinformatics* 2010, **26:**2204-2207.
59. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GR, Ruffier M, Schuster M: **Ensembl 2011.** *Nucleic Acids Res* 2011, **39:**D800-806.
60. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ: **The UCSC Genome Browser database: update 2011.** *Nucleic Acids Res* 2011, **39:**D876-882.
61. Down TA, Piipari M, Hubbard TJ: **Dalliance: interactive genome viewing on the web.** *Bioinformatics* 2011, **27:**889-890.
62. Omitola T, Koumenides CL, Popov IO, Yang Y, Salvadores M, Correndo G, Hall W, Shadbolt N: **Integrating public datasets using linked data: challenges and design principles.** In *Future Internet Assembly: 16-17 December 2010; Ghent, Belgium.* [http://eprints.ecs.soton.ac.uk/21955/]
63. House of Lords, Science and Technology Committee: **Genomic medicine.** London: The Stationery Office Limited; 2009. [http://www.publications.parliament.uk/pa/ld200809/ldselect/ldsctech/107/107i.pdf]

doi:10.1186/gm276
Cite this article as: Dyke SOM, Hubbard TJP: **Developing and implementing an institute-wide data sharing policy.** *Genome Medicine* 2011, **3:**60.