Published in final edited form as:
Nature. 2009 September 10; 461(7261): 168–170. doi:10.1038/461168a.

Benefits and Best Practices of Rapid Pre-Publication Data Release

Toronto 2009 Data Release Workshop Authors

Open discussion of ideas and full disclosure of supporting facts provide the bedrock for scientific discourse and new developments. Traditionally, this has been formally accomplished through published papers, in which both the salient ideas and the supporting facts are combined in a single discrete 'package'. With the advent of methods for large-scale and high-throughput analyses, the generation and transmission of the underlying factual information – the data – are being transformed in an electronic process that involves submitting and retrieving information to and from scientific databases. For most projects, the standard requirement is that all relevant data must be made available at a publicly accessible site at the time of a paper's publication[1].

One of the significant lessons from the Human Genome Project (HGP) was the recognition that making data broadly available prior to publication can be profoundly valuable to the scientific enterprise and lead to public benefits. This is particularly the case when there is a community of scientists that can productively use the data quickly – beyond what the data producers could accomplish themselves in a similar time period, and sometimes for scientific purposes that were not anticipated at the onset of the project. The principles for rapid release of genome-sequence data by the HGP were first formulated at a 1996 meeting held in Bermuda; these were then implemented as policy by several research funding agencies. In exchange for 'early release' of their data, the sequencing groups can request the right to be the first to describe and analyze their complete datasets in peer-reviewed publications. The human genome sequence[2] was the highest profile dataset rapidly released before publication, with assembled sequence data released within 24 hours of generation by each member of the consortium of international sequencing centers. This experience ultimately demonstrated that the broad and early availability of sequence data greatly benefited life sciences research by leading to many new insights and discoveries.

Recognizing that (1) advances in DNA sequencing technologies that allow massive datasets to be produced by an ever-growing number of laboratories have created a need to update policies related to the release of these data, and (2) extending early data release policies to other types of large datasets can be beneficial, a diverse and international group of scientists, ethicists, lawyers, journal editors, and representatives from funding-agencies met in Toronto in May 2009, at a Data Release Workshop convened by Genome Canada and other international agencies. By design, the Toronto meeting continued discussions and policy development planning from previous meetings, in particular: the Bermuda meetings (in 1996, 1997 and 1998), which focused on genome sequence data generated by the HGP[3–5]; the Fort Lauderdale meeting (in 2003), which recommended that rapid pre-publication release be applied to other types of data whose primary utility was a resource for the scientific community, and also established the responsibilities of the resource producers, resource users, and the funding agencies[6]; and the Amsterdam meeting (in 2008), which

Correspondence should be addresses to Ewan Birney (birney@ebi.ac.uk) or Thomas Hudson (tom.hudson@oicr.on.ca).
A complete list of the authors and their affiliations are provided at URL.

extended the scope of rapid data release to proteomics data[7]. Although these meetings' recommendations were applicable to many genomics and proteomics projects, many outside the major centers and funding agencies remain unaware of the details of these policies.

Attendees of the Toronto meeting re-affirmed the value of rapid pre-publication data release for biological and medical datasets that have broad utility and agreed that pre-publication data release should go beyond genomics and proteomics studies to other datasets [e.g., chemical structure, metabolomic, and RNAi datasets, and annotated clinical resources (cohorts, tissue banks, and case-control studies)]. In each of these domains, there are diverse data types and study designs, ranging from large reference projects with broad utility (for which meeting participants endorsed pre-publication data release) to investigator-led hypothesis-testing and data generating projects (for which the minimum standard must be the release of generated data at the time of publication). Several issues discussed at previous data release meetings were not revisited, as they were considered fundamental to all types of data release (whether pre-publication or publication-associated). These included: (1) specification of quality standards for all data; (2) creation of databases designed to facilitate usage of all released data types; (3) archiving of raw data in a retrievable form; (4) housing of both 'finished' and 'unfinished' data in databases; and (5) provision of long-term support for databases by funding agencies. New issues that were addressed include the importance of simultaneously releasing metadata (such as environmental/experimental conditions and phenotypes) that will enable users to fully exploit the data, as well as the complexities associated with human subjects data due to concerns about privacy and confidentiality.

Recommendations for Pre-publication Data Release

At a practical level, the Toronto meeting developed a set of suggested 'best practices' for funding agencies, for scientists in their different roles (e.g., data producers, data analysts/users, and manuscript reviewers), and for journal editors (see Box 1).

Box 1

Guidelines for the Release of Pre-publication Data

1. **Rapid pre-publication data release** should be encouraged for projects with the following attributes:
   - Large-scale (i.e., requiring significant resources over time)
   - Broad utility
   - Creating reference datasets
   - Associated with community buy-in, which is often the case with top-down initiatives
2. **Funding agencies** should facilitate the specification of data release policies for relevant projects by:
   - Explicitly stating any data release requirements, especially mandatory pre-publication data release, in solicitations and instructions to applicants
   - Ensuring that evaluation of data release plans are part of the peer-review process
   - Proactively establishing analysis plans and timelines for projects releasing data pre-publication

    - Fostering investigator-initiated pre-publication data release
    - Helping to develop appropriate consent, security, access and governance mechanisms that protect research participants while encouraging pre-publication data release
    - Providing long-term support of databases

3. **Data producers** should state their intentions and enable analyses of their data by:
    - Informing data users about the data being generated, data standards and quality, planned analyses, timelines, and relevant contact information, ideally through publication of a citeable marker paper near the start of the project or by provision of a citable URL at the project or funding-agency website
    - Providing relevant metadata (e.g., questionnaires, phenotypes, environmental conditions, and laboratory methods) that will assist other researchers in reproducing and/or independently analyzing the data, while protecting interests of individuals enrolled in studies focusing on humans
    - Ensuring that research participants are informed that their data will be shared with other scientists in the research community
    - Publishing their initial global analyses, as stated in the marker paper or citable URL, in a timely fashion
    - Creating databases designed to archive all data (including underlying raw data) in an easily retrievable form and facilitate usage of both pre-processed and processed data
4. **Data analysts/users** should freely analyze released pre-publication data and act responsibly in publishing analyses of those data by:
    - Respecting the scientific etiquette that allows data producers to publish the first global analyses of their dataset
    - Reading the citeable document associated with the project
    - Accurately and completely citing the source of pre-publication data, including the version of the dataset (if appropriate)
    - Being aware that released pre-publication data may be associated with quality issues that will be later rectified by the data producers
    - Contacting the data producers to discuss publication plans in the case of overlap between planned analyses
    - Ensuring that use of data does not harm research participants and is in conformity with ethical approvals
5. **Scientific journal editors** should engage the research community about issues related to pre-publication data release and provide guidance to authors and reviewers on the third-party use of pre-publication data in manuscripts [6]

Funding agencies should require rapid pre-publication data release for projects that generate datasets that have broad utility, are large in scale, and are 'reference' in character. Many such projects have emerged after discussions between funding agencies and the stakeholder scientific community prior to concentrating large amounts of funds in a limited number of data-producing groups, thereby ensuring the efficient generation of the data resource. Table 1 provides examples of projects using different designs, technologies, and approaches that have several of these attributes, but also shows projects that are more hypothesis-based for which pre-publication data release should not be mandated. It was agreed at the meeting that the requirements for pre-publication data release must be made clear when funding opportunities are first announced and that proactive engagement of funders is beneficial throughout the project, as exemplified by the several genome-sequencing projects (e.g., for mouse and many other vertebrates), the International HapMap Project, the ENCODE project, the 1000 Genomes project, and most recently the International Cancer Genome Consortium, the Human Microbiome Project, and the MetaHIT project. For all projects with a data-generation component, the Toronto meeting participants recommended that funding agencies require that data-sharing plans be presented as part of grant applications and that these plans be subjected to peer review. Funding agencies should exercise flexibility in range of circumstances, for example the possibility that large-scale data-generation projects need not necessarily lead to traditional publications, and that certain projects may only need to release some of their generated data prior to publication. Meanwhile, it is desirable to have general consistency in data-sharing policies among funding agencies, whenever possible. At the same time, funding agencies and academic institutions should positively recognize investigators who adopt pre-publication data-release practices; this would be enabled by having released datasets recognized as part of grants and promotion processes as well as tracked using Internet systems similar to those used for traditional publications[8].

Rapid pre-publication data release can lead to tensions between the interests of the data-producing scientists who request a protected time period to publish a first description of a dataset and other scientists who wish to rapidly publish their own analyses based on the same data. To date, many papers have been published by third parties reporting research findings enabled by datasets released prior to publication. These have rarely affected subsequent publications authored by the data producers describing the datasets themselves. Nevertheless, the Toronto meeting participants recognized that this is an ongoing concern that can be addressed by fostering a scientific culture that encourages cooperation on the part of data producers, data analysts, reviewers, and journal editors.

Data producers should, as early as possible and ideally before large-scale data generation begins, clarify their overall plans and intentions for data analysis by providing a citeable statement that can be placed in the publication field of database submissions. This statement must provide clear details about the dataset to be produced, the associated metadata, the experimental design, pilot data, data standards, security, quality control procedures, expected timelines, data release mechanisms, and contact details for lead investigators. If data producers request a protected time period to allow them to be the first to publish the dataset, this should be limited to global analyses of the data and ideally expire within one year. This document would preferably be a ‘marker paper’ that is subjected to peer review and published in a scientific journal. Alternatively, other citeable sources, such as digital object identifiers to specific pages on well-maintained funding agency or institutional web sites, could also be used. Data producers would benefit from defining a citable reference for the database, as it can later be used to reflect impact of the datasets[8].

In turn, the data analysts (i.e., data users) should carefully read the source information associated with a released dataset. Data analysts should pay particular attention to any caveats about data quality, as rapidly released data are not stable, in that they may not have had the full complement of quality control analyses compared to more mature data that become available later in a project. As such, it would be prudent for data analysts to assess the benefits and potential problems in immediately analysing released data. They should communicate with data producers to clarify issues of data quality in relation to the intended analyses, whenever possible. In addition, data users should be aware that some datasets are associated with version numbers: the appropriate version number should be tracked and then provided in any published analyses of those data.

Resulting papers describing studies that do not overlap with the intentions stated by the data producers in the marker paper (or other citeable source) may be submitted for publication at any time, but must appropriately cite the data source. Papers describing studies that do overlap with the data producer's proposed analyses should be handled carefully and respectfully, ideally including a dialogue with the data producer to see if a mutually agreeable publication schedule (such as co-publication or inclusion within a set of companion papers) can be developed. In this regard, it is important for data users to realize that, historically, many such dialogues have led to both coordinated publications and new scientific insights contributed by all parties. Despite the best intentions of all parties, occasional instances might occur when another researcher publishes the results of analyses carried out on pre-publication data and those analyses overlap with the planned studies of the data producer. While such instances are hopefully rare, these should be viewed as a small risk to the data producers, one that comes with the much greater overall benefit of early data release.

As reviewers of manuscripts submitted for publications, scientists should be mindful that pre-publication datasets are likely to have been released before extensive quality control is performed, and any unnoticed errors may cause problems in the analyses performed by third parties. Where the use of pre-publication data is limited or not critical to a study's conclusions, the reviewers should only expect the normal scientific practice of clear citation and interpretation. However, when the main conclusions of a study rely on a pre-publication dataset, reviewers should be satisfied that the quality of the data is described and taken into account in the analysis.

Toronto meeting participants recommended that journals play an active role in the dialogue about rapid pre-publication data release (e.g., in both their guide to authors and instructions to reviewers). Journal editors should encourage reviewers to be aware that large-scale datasets may be subject to specific policies regarding how to cite and use the data. Ultimately, journal editors must rely on their reviewers' recommendations for reaching decisions about publication. By emphasizing the importance of quality review of pre-publication data sets in the manuscript review process, greater awareness and recognition of data producers can be achieved and standards of analysis and publication will be raised.

Data Release for Studies Involving Human Subjects

Clinical, socio-demographic, genomic, and other data about human subjects participating in genetic and epidemiological research studies require particularly careful consideration due to the issues relating to privacy protection and the potential harms that could arise from misuse. These issues are critical to all databases housing information about human subjects, whether or not they contain pre-publication data. These complexities are increased by factors such as managing participant withdrawal or control of data usage once it is in the public domain. For these reasons, it is important to develop and implement robust governance models and procedures for human subjects data early in a project. Lessons can likely be learned from recent models adopted by several projects: *Open Databases* for data variables that cannot be used to identify individuals and *Controlled Access Databases* for clinical and genomic data that are associated with a unique but not directly identifiable individual[9]. Under such conditions, arguments can be made for the release of data for studies involving human subjects, as doing so can augment the opportunities for new discoveries that could ultimately benefit individuals, communities, and society at large.

Conclusion

The rapid pre-publication release of the human genome sequence data by the HGP constituted a landmark model for cooperation between heterogeneous communities of data generators and analysts, successfully demonstrating how 'big science' can be structured for biological research. This data release policy has served the field of genomics well. The benefits of its application to subsequent endeavours have been demonstrated both in providing useful datasets well in advance of a project's completion and in enabling novel scientific advances to be made worldwide. The Toronto meeting participants acknowledged that many issues remain with pre-publication release of data, that there is a range of opinions in the scientific community, that the landscape continues to change rapidly, and policies need to be reviewed on a regular basis. Nonetheless, wider adoption of the general principles that are fundamental to sharing data as early as possible will positively impact the pace of scientific discovery and should be embraced in a practical and well-reasoned fashion.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank the following funding agencies for supporting the Toronto Data Release Workshop: Biotechnology and Biological Sciences Research Council, the European Commission, Genome Canada, the National Human Genome Research Institute, the National Science Foundation, and the Wellcome Trust. We are also grateful to Genny Cardin and the staff at Genome Canada for logistical assistance.

References

1. Cech, TR. Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences. 2003. available at www.nap.edu/books/0309088593/html
2. Lander ES, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 15;409(6822): 860–921.
3. Summary of Principles Agreed at the First International Strategy Meeting on Human Genome Sequencing. 1996. http://www.ornl.org/sci/techresources/Human_Genome/research/bermuda.shtml#1
4. Summary of the Report of the Second International Strategy Meeting on Human Genome Sequencing. 1997. http://www.ornl.org/sci/techresources/Human_Genome/research/bermuda.shtml#2
5. Guyer M. Statement on the rapid release of genomic DNA sequence. Genome Res. 1998 May.8(5): 413. [PubMed: 9582183]
6. Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility. 2003. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf
7. Rodriguez H, et al. Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: The Amsterdam Principles. J Proteome Res. Jul 6; 2009 8(7):3689–3692. [PubMed: 19344107]
8. Credit where credit is overdue. Nature Biotech. Jul.2009 27(7):579. Editorial.
9. Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P. Data sharing in genomics - reshaping scientific practice. Nat Rev Genet. May; 2009 10(5):331–5. [PubMed: 19308065]

Table 1

Examples of pre-publication data release guidelines for different project types.

Project type	**Pre-publication data release recommended**	**Pre-publication data release optional**
Genome sequencing	Whole-genome or mRNA sequence(s) of a reference organism or tissue	Sequences from a few loci for cross-species comparisons in a limited number of samples
Polymorphism discovery	Catalogue of variants from genomic and/or transcriptomic samples in one or more populations	Variants in a gene, a gene family, or a genomic region in selected pedigrees or populations
Genetic association studies	Genome-wide association analysis of thousands of samples	Genotyping of selected gene candidates
Somatic mutation discovery	Catalogue of somatic mutations in exomes or whole- genomes of tumor and non-tumor samples	Somatic mutations of a specific locus or limited set of genomic regions
Microbiome studies	Whole-genome sequence of microbial communities in different environments	Sequencing of target locus in a limited number of microbiome samples
RNA profiling	Whole-genome expression profiles from a large panel of reference samples	Whole-genome expression profiles of a perturbed biological system(s)
Proteomic studies	Mass spectrometry datasets from large panels of normal and disease tissues	Mass spectrometry datasets from a well-defined and limited set of tissues
Metabolomic studies	Catalogue of metabolites in one or more tissues of an organism	Analyses of metabolites induced of a perturbed biological system(s)
RNAi or chemical library screen	Genome-wide screen of a cell line or organism analyzed for standard phenotypes	Focused screens used to validate a hypothetical gene network
3D structure elucidation	Large-scale cataloguing of 3D structures of proteins or compounds	3D structure of a synthetic protein or compound elucidated in the context of a focused project