OPINION

Prepublication data sharing

Rapid release of prepublication data has served the field of genomics well. Attendees at a workshop in Toronto recommend extending the practice to other biological data sets.

Open discussion of ideas and full disclosure of supporting facts are the bedrock for scientific discourse and new developments. Traditionally, published papers combine the salient ideas and the supporting facts in a single discrete 'package'. With the advent of methods for large-scale and high-throughput data analyses, the generation and transmission of the underlying facts are often replaced by an electronic process that involves sending information to and from scientific databases. For such data-intensive projects, the standard requirement is that all relevant data must be made available on a publicly accessible website at the time of a paper's publication[1].

One of the lessons from the Human Genome Project (HGP) was the recognition that making data broadly available before 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 do themselves in a similar time period, and sometimes for scientific purposes outside the original goals of the project.

The principles for rapid release of genome-sequence data from the HGP were formulated at a meeting held in Bermuda in 1996; these were then implemented by several funding agencies. In exchange for 'early release' of their data, the international sequencing centres retained the right to be the first to describe and analyse their complete data sets in peer-reviewed publications. The draft human genome sequence[2] was the highest profile data set rapidly released before publication, with sequence assemblies greater than 1,000 base pairs usually released within 24 hours of generation. This experience demonstrated that the broad and early availability of sequence data greatly benefited life sciences research by leading to many new insights and discoveries[2], including new information on 30 disease genes published prior to the draft sequence.

At a time when advances in DNA sequencing technologies mean that many more laboratories can produce massive data sets, and when an ever-growing number of fields (beyond genome sequencing) are grappling with their own data-sharing policies, a Data Release Workshop was convened in Toronto in May 2009 by Genome Canada and other funding agencies. The meeting brought together a diverse and multinational group of scientists, ethicists, lawyers, journal editors and funding representatives. The goal was to reaffirm and refine, where needed, the policies related to the early release of genomic data, and to extend, if possible, similar data-release policies to other types of large biological data sets — whether from proteomics, biobanking or metabolite research.

Building on the past

By design, the Toronto meeting continued policy discussions from previous meetings, in particular the Bermuda meetings (1996, 1997 and 1998)[3–5] and the 2003 Fort Lauderdale meeting, which recommended that rapid prepublication release be applied to other data sets 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]. A similar 2008 Amsterdam meeting extended the principle of rapid data release to proteomics data[7]. Although the recommendations of these earlier meetings can apply to many genomics and proteomics projects, many outside the major sequencing centres and funding agencies remain unaware of the details of these policies, and so one goal of the Toronto meeting was to reaffirm the existing principles for early data release with a wider group of stakeholders.

In Toronto, attendees endorsed the value of rapid prepublication data release for large reference data sets in biology and medicine that have broad utility and agreed that prepublication data release should go beyond genomics and proteomics studies to other data sets — including chemical structure, metabolomic and RNA interference data sets, and to 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 the large-scale 'community resource projects' first identified at Fort Lauderdale (for which meeting participants endorsed prepublication data release) to investigator-led hypothesis-testing projects (for which the minimum standard should be the release of generated data at the time of publication).

Several issues discussed at previous data-

EXAMPLES OF PREPUBLICATION DATA-RELEASE GUIDELINES

Project type	Prepublication data release recommended	Prepublication 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	Genomewide association analysis of thousands of samples	Genotyping of selected gene candidates
Somatic mutation discovery	Catalogue of somatic mutations in exomes or genomes of tumour and non-tumour 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 data sets from large panels of normal and disease tissues	Mass spectrometry data sets 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 in a perturbed biological system(s)
RNAi or chemical library screen	Large-scale screen of a cell line or organism analysed 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

The Toronto statement

Rapid prepublication data release should be encouraged for projects with the following attributes:
- Large scale (requiring significant resources over time)
- Broad utility
- Creating reference data sets
- Associated with community buy-in

Funding agencies should facilitate the specification of data-release policies for relevant projects by:
- Explicitly informing applicants of data-release requirements, especially mandatory prepublication data release
- Ensuring that evaluation of data release plans is part of the peer-review process
- Proactively establishing analysis plans and timelines for projects releasing data prepublication
- Fostering investigator-initiated prepublication data release
- Helping to develop appropriate consent, security, access and governance mechanisms that protect research participants while encouraging prepublication data release
- Providing long-term support of databases

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 citable 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 analysing 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

Data analysts/users should freely analyse released prepublication 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 data set
- Reading the citeable document associated with the project
- Accurately and completely citing the source of prepublication data, including the version of the data set (if appropriate)
- Being aware that released prepublication 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

Scientific journal editors should engage the research community about issues related to prepublication data release and provide guidance to authors and reviewers on the third-party use of prepublication data in manuscripts

release meetings were not revisited, as they were considered fundamental to all types of data release (whether prepublication or publication-associated). These included: specified quality standards for all data; database designs that meet the needs of both data producers and users alike; archiving of raw data in a retrievable form; housing of both 'finished' and 'unfinished' data in databases; and 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 or experimental conditions and phenotypes) that will enable users to fully exploit the data, as well as the complexities associated with clinical data because of concerns about privacy and confidentiality (see 'Sharing data about human subjects', overleaf).

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

Recommendations for funders

Funding agencies should require rapid prepublication data release for projects that generate data sets that have broad utility, are large in scale, are 'reference' in character and typically have community 'buy-in'. The table opposite provides examples of projects using different designs, technologies, and approaches that have several of these attributes, but also lists projects that are more hypothesis-based for which prepublication data release should not be mandated.

"Funding agencies should require rapid prepublication data release for certain projects."

It was agreed at the meeting that the requirements for prepublication data release must be made clear when funding opportunities are first announced and that proactive engagement of funders is beneficial throughout a project, as has been the experience of many genome-sequencing efforts, the International HapMap Project, the ENCODE project, the 1000 Genomes project and, more recently, the International Cancer Genome Consortium, the Human Microbiome Project and the MetaHIT project.

For all projects generating large data sets, the Toronto meeting recommended that funding agencies require that data-sharing plans be presented as part of grant applications and that these plans are subjected to peer review. Such practice is currently the exception rather than the rule. Funding agencies will need to exercise flexibility by, for example, recognizing that large-scale data-generation projects need not necessarily lead to traditional publications, and that certain projects may need to release only some of their generated data before publication. At the same time, general consistency in data-sharing policies between funding agencies is desirable, whenever possible. To encourage compliance, funding agencies and academic institutions should give credit to investigators who adopt prepublication data-release practices, one option would be to recognize good data-release behaviour during grant renewals and promotion processes, another would be to track the usage and citation of data sets using electronic systems similar to those used for traditional publications[8].

Data producers and data users

Early data release can lead to tensions between the interests of the data-producing scientists who request the right to publish a first description of a data set and other scientists who wish to publish their own analyses of the same data. To date, many papers have been published by third parties reporting research findings enabled by data sets released before publication. The experiences shared in Toronto suggest that these have rarely affected subsequent publications authored by the data producers. Nevertheless, the Toronto meeting participants recognized that this is an ongoing concern that is best addressed by fostering a scientific culture that encourages transparent and explicit 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 intentions for data analysis by providing a citable statement, typically a 'marker paper', that would be associated with their database entries. This statement should provide clear details about the data set 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 data set, this should be limited to global analyses of the data and ideally expire within one year.

If the citable statement is a 'marker paper' it should be subjected to peer review and published in a scientific journal. Alternatively, other citable sources, such as digital object identifiers to specific pages on well-maintained funding agency or institutional websites, could also be used. Data producers benefit from creating a citable reference, as it can later be used to reflect impact of the data sets[8].

In turn, the data users should carefully read the source information, including any marker papers, associated with a released data set. Data analysts should pay particular attention to any caveats about data quality, because rapidly released data are often unstable, in that they may not yet have been subjected to full quality control and so may change. 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 data sets are associated with version numbers: the appropriate version number should be tracked and then provided in any published analyses of those data.

> "Prepublication data are likely to be released before extensive quality control is performed."

Resulting papers describing studies that do not overlap with the intentions stated by the data producers in the marker paper (or other citable 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 coordinated publications and to new scientific insights. Despite the best intentions of all parties, on occasion a researcher may publish the results of analyses that overlap with the planned studies of the data producer. Although such instances are hopefully rare if good communication protocols are followed, 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.

Editors and reviewers

As reviewers of manuscripts submitted for publication, scientists should be mindful that prepublication data sets 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 prepublication data is limited or not crucial 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 prepublication data set, reviewers should be satisfied that the quality of the data is described and taken into account in the analysis.

Participants at the Toronto meeting recommended that journals play an active part in the dialogue about rapid prepublication data release (both in their formal guide to authors and informal instructions to reviewers). Journal editors should remind reviewers that large-scale data sets 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. However, encouraging reviewers to carefully check the conditions for using data that authors have not created themselves can help to raise both the quality of analysis and fairness in citation of published studies.

Conclusion

The rapid prepublication release of sequencing data has served the field of genomics well. The Toronto meeting participants acknowledged that policies for prepublication release of data need to evolve with the changing research landscape, that there is a range of opinion in the scientific community, and that actual community behaviour (as opposed to intentions) need to be reviewed on a regular basis. To this end, we encourage readers to join the debate over data-sharing principles and practice in an online forum hosted at http://tinyurl.com/lqxpg3. ■

Toronto International Data Release Workshop Authors. A complete list of the authors and their affiliations accompanies this article online.
e-mail: birney@ebi.ac.uk, tom.hudson@oicr.on.ca

Sharing data about human subjects

Data about human subjects participating in genetic and epidemiological research require particularly careful consideration owing to privacy-protection issues 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 prepublication data.

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 probably be learned from data policies adopted by several genomics projects[9] that generate human-subject data. For aggregated data that cannot be used to identify individuals, databases are open access, but for clinical and genomic data that are associated with a unique, but not directly identifiable individual, access may be restricted.

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.

1. Committee on Responsibilities of Authorship in the Biological Sciences, National Research Council *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences* (National Academy of Sciences, 2003).
2. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
3. *Summary of Principles Agreed at the First International Strategy Meeting on Human Genome Sequencing* Bermuda, 25–28 February 1996 (HUGO, 1996); available at www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml
4. Summary of the Report of the Second International Strategy Meeting on Human Genome Sequencing Bermuda, 27 February–2 March 1997 (HUGO, 1997); available at www.ornl.org/sci/techresources/Human_Genome/research/bermuda.shtml#2
5. Guyer, M. *Genome Res.* **8,** 413 (1998).
6. *Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility* (Wellcome Trust, 2003); available at www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf
7. Rodriguez, H. *et al. J. Proteome Res.* **8,** 3689–3692 (2009).
8. *Nature Biotechnol.* **27,** 579 (2009).
9. Kaye, J., Heeney, C., Hawkins, N., de Vries, J. & Boddington, P. *Nature Rev. Genet.* **10,** 331–335 (2009).

Join the discussion at http://tinyurl.com/lqxpg3

See online special at http://tinyurl.com/dataspecial