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.

The Wellcome Trust

Introduction

The Wellcome Trust sponsored a meeting on 14–15 January 2003 to discuss how, at this point in the development of the field of genomics, pre-publication data release can promote the best interests of science and help to maximize the public benefit to be gained from research. About 40 people attended the meeting, among them large-scale sequence producers, sequence users including computational biologists, representatives of the major nucleotide sequence databases, journal editors, and scientists interested in other large-scale data sets. The discussion took as a given that published data are available in their entirety for any use by any investigator, and focused on issues involved in making data broadly available *prior* to publication.

The meeting concluded that pre-publication release of sequence data by the International Human Genome Sequencing Consortium, and other sequence producers, has been of tremendous benefit to the scientific research community in general. While not all were in a position to make commitments for their funding agencies, the meeting attendees were in broad agreement that, to encourage the continuation of such benefits, the sequence producers, sequence users and the funding agencies recognize and implement a system based on 'tripartite responsibility'. Specifically,

- The meeting attendees enthusiastically reaffirmed the 1996 Bermuda Principles, which expressly called for rapid release to the public international DNA sequence databases (GenBank, EMBL, and DDBJ) of sequence assemblies of 2kb or greater by large-scale sequencing efforts and recommended that that agreement be extended to apply to all sequence data, including both the raw traces submitted to the Trace Repositories at NCBI and Ensembl and whole genome shotgun assemblies.
- The attendees recommended that the principle of rapid pre-publication release should apply to other types of data from other large-scale production centers specifically established as 'community resource projects'.
- The attendees recognized that pre-publication data release might conflict with a fundamental scientific incentive – publishing the first analysis of one's own data. The attendees noted that it would not be possible to absolutely guarantee this incentive without applying restrictions that would undermine the rationale for rapid, unrestricted release of data from community resources. Nonetheless, it is essential that excellent scientists continue to be attracted to these projects. To encourage this, the scientific community should understand that pre-publication data release needs active community-wide support if it is to continue to receive widespread support from the producers. The contributions and interests of the large-scale data producers should be recognized and respected by the users of the data, and the ability of the production centres to analyse and publish their own data should be supported by their funding agencies.

Community resource projects

A 'community resource project' is a research project specifically devised and implemented to create a set of data, reagents or other material whose primary utility will be as a resource for the broad scientific community. Recent examples of community resource projects include the International Human Genome Sequencing Consortium, the Mouse Genome Sequencing Consortium, the Mammalian Gene Collection, the SNPs Consortium, and the International

HapMap Project. The products of community resource projects have, over the past several years, become increasingly important as drivers of progress in biomedical research. The scientific community will best be served if the results of community resource projects are made immediately available for free and unrestricted use by the scientific community to engage in the full range of opportunities for creative science. At the same time, it is crucial that the scientific community recognizes and respects the important contribution made by the scientists who carry out community resource projects.

Tripartite sharing of responsibility

An optimized system for generating community resources involves three constituencies within the scientific research community – resource producers, resource users, and funding agencies. Each of the three has a unique and critical role to play in ensuring the growth and development of the community resource system.

A. **Funding agencies**. Funding agencies are the major sources of support of research projects leading to community resources and projects that depend on the availability of such resources. Funding agencies have a critical role in determining the quality and breadth of community resources through the peer review evaluation system and as the sources of scientific research policies. For these reasons funding agencies should:

   1. designate appropriate efforts as community resource projects, and encourage resource producers to prepare and submit Project Descriptions (see below) for publication;
   2. require, as a condition of funding, free and unrestricted data release from community resource projects to appropriate central and searchable public databases, and vigorously ensure that this occurs;
   3. encourage more investigators to serve the community through involvement in such projects. In particular, the agencies should ensure that investigators engaged in generation of such datasets have sufficient support for curation, maintenance and distribution of the data to the community, as well as resources to perform initial analyses using the resources that they have generated;
   4. ensure that a centralized view of existing community resource projects is available as an information source for the community;
   5. support central databases that will house and distribute the data in a way that prevents fragmentation of the data.

B. **Resource producers**. Community resources are often expensive efforts. For this and other reasons, they are frequently established and supported as unique facilities. The scientists who organize and operate community resources are, accordingly, in a uniquely responsible position. The community is dependent on the success of their efforts and they often face relatively little direct competition. Resource producers should:

   1. when feasible, publish a Project Description. The purpose of the Project Description, which will be a new type of scientific publication, is to inform the scientific community about the resource project and to provide a citation to reference the source of the data. The Project Description should be written at the beginning of the project and describes

the plans for and scope of the production and analyses that the data producer intends to undertake. It will often include a timeline for production goals and data release.

2. produce data of consistently high quality;
3. make the data generated by the resource immediately and freely available without restriction;
4. recognize that even if the resource is occasionally used in ways that violate normal standards of scientific etiquette, this is a necessary risk set against the considerable benefits of immediate data release.

C. **Resource Users.** Community resource data sets benefit the users enormously, giving them the opportunity to analyse the data without the need to generate it first. The data sets are, in general, much larger, richer and of higher quality than individual laboratories could normally generate. In contributing to what ideally is a symbiotic and synergistic situation, resource users should:

1. appropriately cite the source of the data analysed and acknowledge the resource producers. The early publication of a Project Description, as suggested above, would provide users with an appropriate reference to cite before the data are formally published;
2. recognize that the resource producers have a legitimate interest in publishing prominent peer-reviewed reports describing and analyzing the resource that they have produced (and that neither the Project Descriptions nor data deposits in databases are the equivalent of such publications);
3. respect the producer's legitimate interests as set out, e.g. in a Project Description, while being free to use the data in any creative way. There should be no restrictions on the use of the data, but the best interests of the community are served when all act responsibly to promote the highest standards of respect for the scientific contribution of others. In some cases, this might best be done by discussion or coordination with the resource producers;
4. assist journals and funding agencies to play their proper roles in ensuring, through the peer review system, that the system works fairly for all constituents.

Large-scale genome sequencing

Large-scale genome sequencing projects are clearly community resource projects, and serve as a well-developed example to illustrate the general principles described above. The Bermuda Principles (http://www.gene.ucl.ac.uk/hugo/bermuda.htm) were developed in 1996 by the scientists engaged in the International Human Genome Sequencing Consortium and their funding agencies, and have been the basis of a successful system for achieving rapid and open data release. Now, in 2003, the meeting attendees, recognizing the role of users as well as producers and funders in effecting a successful system, enthusiastically recommend the reaffirmation of the Bermuda Principles for continued large-scale sequencing projects, and recommend that:

1. They should be extended beyond their initial application to sequence assemblies of a minimum size from BAC-based sequence projects so that they apply to rapid (i.e. as soon as possible) release of both raw and assembled sequence data, subject only to the data meeting appropriate quality assessment standards.

2. Funding agencies, users and sequencing centers should all honor their obligations, as described above.

Other community resource projects

In the near future, many other large data sets will be produced as community resources. While only a few of the meeting attendees were familiar with data types other than large-scale sequences, the attendees recommended that appropriate implementation of the principles discussed should be devised for other community resource projects, such as large-scale protein structure determination or gene expression analysis. In many of these cases, the solutions, in terms of such considerations as data quality standards, data storage and dissemination modes, and producer and user interests, are only beginning to emerge. Development of effective systems for achieving the objectives of the community resource concept should be an integral component of the planning and development of such new community resources.

Research materials and tools

Some of the issues involved in ensuring rapid and open access to finite resources, such as reagents, clones, cell lines and other material resources, are different from those pertaining to electronic data sets. The meeting attendees strongly encouraged the relevant funding agencies, resource producers and users to develop practical approaches to maximizing the benefit of this type of resource to the scientific community and to research.

Pre-publication release of other data

Beyond community resource projects, many valuable data sets could come from other sources. Still different issues arise in the case of resources that emerge from research efforts whose primary goal is not resource generation. In such cases, contribution of the data to the public domain as a resource is more a voluntary matter. To obtain the clear benefit that would ensue from converting such data sets into community resources as rapidly as possible, incentives should be developed by the scientific community to support the voluntary release of such data prior to publication, by appropriately recognizing and protecting the interests of scientists who wish to share such pre-publication data with the community.

The Wellcome Trust is an independent research-funding charity, established under the will of Sir Henry Wellcome in 1936. It is funded from a private endowment, which is managed with long-term stability and growth in mind.

Its mission is to foster and promote research with the aim of improving human and animal health. Its work covers four areas:

Knowledge – improving our understanding of human and animal biology in health and disease, and of the past and present role of medicine in society.

Resources – providing exceptional researchers with the infrastructural and career support they need to fulfil their potential.

Translation – ensuring maximum health benefits are gained from biomedical research.

Public engagement – raising awareness of the medical, ethical and social implications of biomedical science.

The Wellcome Trust
183 Euston Road
London NW1 2BE
Tel: +44 (0)20 7611 8888
Fax: +44 (0)20 7611 8545
E-mail: contact@wellcome.ac.uk
Web: www.wellcome.ac.uk