Information- and Data-Sharing Policy

Genomic Science Program

Office of Biological and Environmental Research

Office of Science

Department of Energy

Draft Date: May 9, 2013

Introduction

The changing scope of scientific inquiry and the astonishing rate of data production in systems biology research drive the development of new types of data-centric cyber infrastructures, which, in turn promotes and fosters data and information sharing among researchers. Journals, funding agencies, and governments correspondingly have developed information standards and sharing policies—all of which, in one way or another, address conducting research in an open-access environment (1, 2, 3). A key hallmark of data- and information-sharing policies is the requirement that scientific inquiry and publications must include the submission of publication-relevant information and materials into public repositories. For the most part, data-sharing policies follow the uniform principle for sharing integral data and materials expeditiously (called UPSIDE) (4). Conversely, when research information is not made publicly available to a global scientific community, a corresponding price is paid in lost opportunities, barriers to innovation and collaboration, and the obvious problem of unknowing repetition of similar work (5).

This statement summarizes the data- and information-sharing policy of the Genomic Science research program at the Department of Energy's (DOE) Office of Biological and Environmental Research (BER). BER recognizes that tackling challenging scientific questions requires research data to be made publically accessible and in formats that encourage reuse and integration for greater interoperability. We affirm our support for the concept of information and data sharing in order to encourage researchers to exchange new ideas, data, and technologies across the Genomic Science program and the wider scientific community.

The DOE Office of Science will require that all new, renewal, and supplemental applications develop a digital data-management plan as part of a full proposal submission. The Genomic Science program data- and information-sharing policy is supplemental to the Office of Science policy and requires that all publication-related information and data be made publically available.

Policy Statement:

The Biological and Environmental Research (BER) Program requires that all publishable data, metadata, and software resulting from research funded by the Genomic Science program must conform to community-recognized standard formats when they exist, be clearly attributable, and be deposited within a community-recognized public database(s) appropriate for the research. Publication-related data that are consistent with the data and analysis models of the DOE Systems Biology Knowledgebase (KBase) should be accessible to KBase and could be integrated into KBase. All digitally accessible data obtained as a result of research funded by the Genomic Science program must be kept in an archive maintained by the Principal Investigator (PI) for the duration of the funded project. Any publications resulting from the use of shared data must accurately acknowledge the original source or provider of the attributable data. The publication of information resulting from research funded by the Genomic Science program must be consistent with the intellectual property provisions of the contract under which the publishable information was produced.

I. Applicability

This policy shall apply to all projects receiving funding in the Genomic Science program as of October 1, 2013, which exceeds the requirement of the Office of Science policy that only applies to solicitations issued on or after October 1, 2013. For cases where information-sharing standards or databases do not yet exist, the information-sharing and data-archiving plan provided by a project's PI must state these limitations. Data and information that are necessary elements of protected intellectual property and related to a pending or future patent application are explicitly exempt from public access until completion of the patenting process. Adherence to this policy will be monitored through the established procedure of yearly progress reports submitted to Genomic Science program managers. All information about the Genomic Science data-sharing policy will be made publicly available at http://genomicscience.energy.gov/datasharing/.

This policy supports the definition of **digital research data** as defined by the DOE Office of Science. Research data is the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This 'recorded' material excludes physical objects (e.g., laboratory samples).

II. Submission of Publication-Related Information and Data

All investigators funded by the Genomic Science program are expected to submit their publication-related data and information to a national or international public repository, when one exists, according to the repository's established standards for content and timeliness; (when no public repository exists see Section II.B.2, "Other Technologies" for policy). This includes:

- Experimental protocols and/or workflows
- Raw and/or processed data, as required by the repository
- Other relevant supporting materials and/or metadata.

BER will maintain a website listing all peer-reviewed and published papers and issued patents resulting from Genomic Science program funding. PIs are expected to inform BER when a publication or patent appears in print or online. BER recognizes that sub-disciplines and experimental technologies have varying degrees of cyber-infrastructure and standard ontology to accommodate this policy. Specific guidelines and suggestions for Genomic Science investigators are provided below.

II.A Nationally and Internationally Accepted Databases and Ontologies

IIA.1 DOE Systems Biology Knowledgebase

A long-term vision for the Genomic Science program is to develop an integrated data and modeling computational environment for systems biology (6). In support of this vision, in 2011 BER launched the DOE System Biology Knowledgebase (KBase, KBase.us). KBase is an *open-source and open-architecture* computational environment for integrating large, diverse datasets generated by the scientific community and for using this information to advance predictive understanding, manipulation, and design of biological processes in an environmental context. KBase is meant to be a community resource that enables users to integrate a wide spectrum of genomics and systems biology data, models, and information for microbes, microbial communities, and plants. Powerful tools available within KBase can analyze and simulate data to predict biological behavior, generate and test hypotheses, design new biological functions, and propose new experiments.

KBase provides predictive biology capabilities based in part on a common data model that is expected to be extended over time to support more comprehensive and accurate models of biological systems. To achieve high-quality predictions, it is imperative that the data and associated metadata be also high-quality. Currently, the supported high-level data types include genomes (bacteria, archea, eukaryotes), transcriptomes, phenotypes, 16s amplicons, metagenomes, proteomes (mapped to genomes), variation, and interactomes. KBase will be pulling these data types from existing international repositories (see https://www.kbase.us/about/kbase-data-policy/kbase-data-sources/) for list of such repositories). Thus, data submitted to these standard resources will ultimately be integrated into the KBase system. It is possible to submit certain forms of the above data directly to KBase (see https://www.kbase.us/about/kbase-data-policy/kbase-data-types/), although integration into KBase must pass minimal quality and metadata standards before acceptance.

II.A.2 Sequence Data

The field of genomic sequencing has a well-developed mechanism for public archiving of experimental data. Nucleotide sequence data will be deposited into National Center for Biotechnology Information (NCBI) GenBank, and protein sequence data will be deposited into one of the community Protein Sequence Databases such as UniProtkb/Swiss-Prot Protein Knowledge database. In addition, investigators are encouraged to use the Gene Ontology Annotation Database (7) when possible and to follow the guidelines on minimum information standards provided by the Genomic Standards Consortium (GSC).

Specifically for large-scale sequencing projects funded by the Genomic Science program, whole-genome sequencing data must be made publicly available after first assembly of the sequencing reads for that genome. In the case of metagenomic sequencing, data must be deposited to the National Center for Biotechnology Information (NCBI) after completion of the last sequencing run. For research carried out by the DOE Joint Genome Institute, DOE-funded investigators will follow the JGI data release policy (http://my.jgi.doe.gov/general/datarelease.html).

II.A.3 Three-Dimensional Structural Data

All coordinates and related information for structures of biological macromolecules and complexes are to be deposited in the Protein Data Bank (PDB) or Nucleic Acid Database (NDB), as defined by the PDB and NDB.

II.A.4 Gene Expression Data

The Microarray Gene Expression Data Society (MGED, www.mged.org; now known as Functional Genomics Data Society (FGED), www.fged.org) recommends the use of MGED ontology for the description of key experimental conditions as, for example, using a MIAME or MINSEQE compliant format. BER's policy follows the MGED-recommended ontology. Genomic Science program researchers are strongly encouraged to deposit raw and transformed data sets and experimental protocols and/or workflows to a public gene expression database. Possible gene expression databases for data deposition include the Gene Expression Omnibus (8) and ArrayExpress (9).

IIA.5 Standard Identification Numbers and Names

BER further adopts and encourages the use of standard identification numbers, such as Enzyme Commission numbers (EC), Digital Object Identifiers (DOIs), and ICSB bacterial names.

II.B Information-Sharing Systems and Databases Under Development

II.B.1 Proteomics

The Proteomics Standards Initiative (PSI), a working group of the Human Proteome Organization (HUPO, www.hupo.org), has outlined a number of standards initiatives to represent molecular interactions (MIMIx), mass spectrometry (MIAPE_MS), proteomics informatics (MIAPE_MSI), and protein separations (MIAPE_GE). Because this is an evolving initiative and the field is still immature, we encourage researchers to engage with HUPO and use these standards when appropriate. Public proteomics-related repository databases are being developed, such as the Open Proteomic Database (10), PRIDE (11), and PeptideAtlas (12). Genomic Science program researchers are encouraged to engage with these databases. However, standards and ontologies will evolve within the proteomics community, and the Genomic Science program data policy will follow guidelines set forth by HUPO as they develop.

II.B.2 Other Technologies

Genomic Science research makes use of a large variety of technologies for which there are, as yet, no national or international information standards and archival formats. Scientists in the Genomic Science program are encouraged to participate in the efforts of research communities to develop new standards for enabling information sharing.

It is the long term objective of BER to encourage the development of infrastructure for technologies that do not as yet have nationally or internationally accepted information-sharing standards. In cases where there are no public repositories or community-driven standard ontologies, BER encourages the PI to make these types of data and information publicly available whenever possible and by the most appropriate available method as determined by the PI.

III. Protection of Human Subjects

Research using human subjects provides important scientific benefits, but these benefits never outweigh the need to protect individual rights and interests. BER requires that grantees and contractors follow the DOE principles and regulations for the protection of human subjects involved in DOE research. Minimally, this will require an Institutional Review Board (IRB) review. These principles are stated clearly in the Policy and Order documents: DOE P 443.1A and DOE O 443.1A, which are available online at www.directives.doe.gov and http://humansubjects.energy.gov/.

IV. Computational Software

The International Society for Computational Biology (ISCB) recommends that funding agencies follow ISCB guidelines for open-source software development. ISCB states that proposals for developing research software should be specific about cost, source code availability, redistribution rights (including for derived works), user support, and any discrimination among user types. BER requires that research software developed with Genomic Science program funding and that result in a peer-reviewed software publication be made accessible through either an open-source license (www.opensource.org) or be deposited in an open source software community, such as SourceForge or GitHub.

V. Laboratory Data Management and Archiving

Systems biology research projects that involve high-throughput, data-intensive research may necessitate the use of a data management system to automatically handle this pipeline of data. BER’s goal is that researchers within the Genomic Science program develop a laboratory data management plan or system for managing high-throughput research data and information. Different research projects require different information management systems. Research projects that involve more than one senior investigator will be required to implement a laboratory data management system or a similar type of electronic system for data and information sharing, archiving, and retrieval. Smaller projects are also strongly encouraged to develop data management approaches or systems.

VII. Summary

This document outlines the Genomic Science program data- and information-sharing policy. BER (and the Office of Science) requires that all Genomic Science program-funded principal investigators construct a Data Management Plan as a component of their research proposals. The policy requires data to conform to existing community-recognized standard formats whenever possible, to be clearly attributable, and to be deposited in a timely manner to a community-recognized public database(s) appropriate for the research conducted. BER is committed to encouraging development of data and information repositories and standard ontologies. BER recognizes that this policy will necessarily be revised to include new standards, new data types, and other advances that are pertinent to maximizing availability of data and information across the Genomic Science program. This information- and data-sharing policy and related materials can be found at http://genomicscience.energy.gov/datasharing/.

References

1. Field, D., et al. 2009. "'Omics Data Sharing," *Science* **326,** 234–236.
2. http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research
3. http://www.whitehouse.gov/sites/default/files/omb/assets/omb/circulars/a110/2cfr215-0.pdf
4. National Research Council. 2003. *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. National Academy of Sciences. http://www.nap.edu/openbook.php?isbn=0309088593.
5. Uhlir, P. F., and P. Schröder. 2007. "Open Data for Global Science," *Data Science Journal* **6**, OD36–OD53.
6. U.S. DOE. 2005. *Genomics:GTL Roadmap: Systems Biology for Energy and Environment*, DOE/SC-0090. U.S. Department of Energy Office of Science. http://genomicscience.energy.gov/roadmap/.
7. Camon, E., et al. 2004. "The Gene Ontology Annotation (GOA) Database: Sharing Knowledge in Uniprot with Gene Ontology," *Nucleic Acids Res*. **32**, D262–D266.
8. Edgar, R., M. Domrachev, and A. E. Lash. 2002. "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository," *Nucleic Acids Res*. **30**, 207–210.
9. Brazma, A., et al. 2003. "ArrayExpress—A Public Repository for Microarray Gene Expression Data at the EBI," *Nucleic Acids Res*. **31**, 68–71.
10. Prince, J. T., et al. 2004. "The Need for a Public Proteomics Repository," *Nature Biotechnology*, **22**, 471–472.

11. Martens, L., et al. 2005. “PRIDE: the Proteomics Identifications Database,” *Proteomics,* **5**, 3537–3545.
12. Deutsch, E. W., H. Lam, and R. Aebersold. 2008. “PeptideAtlas: A Resource for Target Selection for Emerging Targeted Proteomics Workflows,” *EMBO Rep*., **9**, 429–424.