Data Protection & Release Guidelines

Contents

Data Protection & Release Guidelines

DATA SHARING

It is the policy of Genome BC that data produced with its funds and support will be freely shared and made widely available, thereby enabling the effortless stream of data within and between fields. Data will be shared in a manner adhering to applicable laws, the policies of other project funders and with full and proper attribution to the data provider. Genome BC recognizes that in some instances the data will include propriety information and/or know-how which may be desirable to protect, in a reasonable period of time, prior to mandatory public dissemination of the data.

GOALS OF DATA SHARING

Data sharing and data documentation allow scientists to facilitate the translation of research results into knowledge, products, procedures and measurable outcomes related to Genome BC's contracts and initiatives.

There are many reasons to share data. Sharing data facilitates open scientific inquiry, encourages diversity of analysis and opinion; promotes new research; makes possible the testing of new or alternative hypotheses and methods of analysis; supports studies on data collection methods and measurement; facilitates the education of new researchers; enables the exploration of topics not envisioned by the initial investigators; permits the creation of new datasets for new applications when data from multiple sources are combined; and enables open inquiry about the impact of project activities.

In short, *data sharing maximizes the value of each project's outputs.* Genome BC-funded projects must follow the data release and resource sharing principles of a "community resource project", defined as 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.

APPLICABILITY OF GENOME BC DATA PROTECTION & RELEASE GUIDELINES

The Genome BC Data Protection & Release Guidelines apply to:

- the sharing of research data for not-for-profit research and charitable purposes,
- final research data, and the basic research, laboratory studies, field surveys, and other types of research supported by Genome BC (Note: It is especially important to share unique data that cannot be readily replicated), and
- projects that transform or link pre-existing datasets (as opposed to producing new data). (Note: If there are limitations associated with a data sharing agreement of the original data that preclude subsequent sharing, then the principal investigator should explain this in the project proposal.)

Genome BC also understands that data distribution regulations vary by country. Investigators collecting data in non-Canadian countries should familiarize themselves

with the policies and laws governing data sharing in the country(s) in which they plan to work and address specific limitations in their data-sharing strategy.

Even if Genome BC support is sought to transform or link datasets (as opposed to producing a new set of data), the investigator should still include a data-sharing agreement for the original data that preclude subsequent sharing, then the applicant should explain this in the application.

IMPLEMENTATION

Given the variety of projects that Genome BC supports, neither the precise content for the data documentation, nor the formatting, presentation, or transport mode for data is stipulated. What works for one field or one study may not work for others.

However, principal investigators must plan for data sharing, and be aware of the current state of data sharing activities and data-management best practices within their disciplines and fields. This means the principal investigator is expected to participate in and contribute to the creation of environments that support data sharing by seeking out relevant data sharing activities and networking with others within their discipline. These activities address areas such as:

- relevant online data repositories (e.g., Genbank) and data federations,
- procedures for data documentation,
- data formatting and data exchange standards,
- software (or online data services) that conform to data format and exchange standards,
- procedures for quantifying the demand (use) for the data (i.e., number and rate of users and records per data repository and/or data federation per year), and
- procedures for the data owner, or provider of the datasets, to receive feedback about the quality of the data.

Familiarity with these subjects will allow the principal investigator to better estimate the costs and benefits associated with their data sharing strategy and implementation plan.

TIMELINESS OF DATA SHARING

Recognizing that the value of data often depends on their timeliness, data sharing should occur in a timely fashion. Genome BC expects the timely release and sharing of data to be no later than the acceptance for publication of the main findings from the final dataset. The specific time will be influenced by the nature of the data collected. Data from small studies can be analyzed and submitted for publication relatively quickly. If data from large epidemiologic or longitudinal studies are collected over several discrete time periods or waves, it is reasonable to expect that the data would be released in waves as data become available or main findings from waves of the data are published. Genome BC recognizes that the investigators who collected the data have a legitimate interest in benefiting from their investment of time and effort. Genome BC continues to expect that the initial investigators may benefit from first and continuing use but not necessarily from prolonged exclusive use.

HUMAN SUBJECTS AND PRIVACY ISSUES

The rights and privacy of human subjects who participate in Genome BC-sponsored research must be protected at all times. It is the responsibility of the investigators, their Institutional Review Board (IRB), and their institution to protect the rights of subjects and the confidentiality of the data. Prior to sharing, data should be redacted to strip all identifiers, and effective strategies should be adopted to minimize risks of unauthorized disclosure of personal identifiers. Stripping a dataset of items that could identify individual participants is referred to by several different terms, such as "data redaction," "de-identification of data," and anonymizing data. In addition to removing direct identifiers, e.g., name, address, telephone numbers, and social insurance numbers, researchers should consider removing indirect identifiers and other information that could lead to "deductive disclosure" of participants' identities. Deductive disclosure of individual subjects becomes more likely when there are unusual characteristics of the joint occurrence of several unusual variables. Samples drawn from small geographic areas, rare populations, and linked datasets can present particular challenges to the protection of subjects' identities.

Also, if research participants are promised that their data will not be shared with other researchers, the application should explain the reasons for such promises. Such promises should not be made routinely and without adequate justification. For the most part, it is not appropriate for the initial investigator to place limits on the research questions or methods other investigators might pursue with the data. It is also not appropriate for the investigator who produced the data to require co-authorship as a condition for sharing the data.

Investigators may use different methods to reduce the risk of subject identification. One possible approach is to withhold some part of the data. Another approach is to statistically alter the data in ways that will not compromise secondary analyses but will protect individual subjects' identities. Alternatively, an investigator may restrict access to the data at a controlled site, sometimes referred to as a data enclave. Some investigators may employ hybrid methods, such as releasing a highly redacted dataset for general use but providing access to more sensitive data with stricter controls through a data enclave.

Researchers who seek access to individual level data are typically required to enter into a data-sharing agreement. Data-sharing agreements, which come by many terms, including "license agreements," and "data distribution agreements," generally include requirements to protect participants' privacy and data confidentiality. They may prohibit the recipient from transferring the data to other users or require that the data be used for research purposes only, among other provisions, and they may stipulate penalties for violations. For further information on these alternative mechanisms to share data while protecting participant confidentiality, see also the section concerning "Methods for Data Sharing." In most instances, sharing and archiving of data is possible without compromising confidentiality and privacy rights. The procedures adopted to share data while protecting privacy should be individually tailored to the specific dataset.

Investigators seeking Genome BC support for clinical trials may wish to consider several factors as they develop their data-sharing plan. Researchers who are planning clinical trials and intend to share the resulting data should think carefully about the study design, the informed consent documents, and the structure of the resulting

dataset prior to the initiation of the study. For example, many early phase clinical trials use small samples, which make it difficult to protect the privacy of the participants. Furthermore, some study designs afford greater privacy protection to subjects than others. For example, longitudinal research poses challenges because the need to retain identifiers in order to link individual-specific data collected at different time points.

Many research efforts supported by Genome BC do not include human subjects. Final research datasets from studies that do not include human subjects generally should not be constrained by the limitations deemed necessary and appropriate for human subjects.

PROPRIETARY DATA

Issues related to proprietary data also can arise with corresponding constraints on public disclosure. Genome BC recognizes the need to protect patentable and other proprietary data. Any restrictions on data sharing due to a need for protection of arising data and ensuing discoveries should be discussed in the data-sharing plan section of an application and will be considered by program staff. While Genome BC understands that an institution's desire to exercise its intellectual property rights may justify a need to delay disclosure of research findings, a delay of 30 to 90 days is generally viewed as a reasonable period for such activity, unless there are extenuating circumstances that are openly disclosed to and agreed upon by all parties concerned.

METHODS FOR DATA SHARING

There are many ways to share data.

- Publishing
- Patents
- Researcher's publicly and freely accessible sources (e.g. website)
- Data Archives (e.g. GenBank, EMBL, etc.)
- Open Source Archives (sourceforge.net)
- Depositing into Strain Collections (e.g. ATCC)

When making data available, researchers cannot place limits on questions posed, methods used, nor require co-authorship as a condition for receiving data. Although we recognize that certain end users will have special needs.

There are several mechanisms for data sharing that investigators can use. For example, investigators sharing under their own auspices should consider using a **data-sharing agreement** to impose appropriate limitations on users. Such an agreement usually indicates the criteria for data access, whether or not there are any conditions for research use, and can incorporate privacy and confidentiality standards to ensure data security at the recipient site and prohibit manipulation of data for the purposes of identifying subjects.

The method for sharing that an investigator selects is likely to depend on several factors, including the sensitivity of the data, the size and complexity of the dataset, and the volume of requests anticipated. Investigators sharing under their own auspices may simply mail a CD with the data to the requestor, or post the data on their

institutional or personal website. Although not a condition for data access, some investigators sharing under their own auspices may form collaborations with other investigators seeking their data in order to pursue research of mutual interest. Others may simply share the data by transferring them to a data archive facility to distribute more widely to interested users, to maintain associated documentation, and to meet reporting requirements. Data archives can be particularly attractive for investigators concerned about a large volume of requests, vetting frivolous or inappropriate requests, or providing technical assistance for users seeking help with analyses.

Datasets that cannot be distributed to the general public, for example, because of participant confidentiality concerns, third-party licensing or use agreements that prohibit redistribution, or national security considerations, can be accessed through a data enclave. A data enclave provides a controlled, secure environment in which eligible researchers can perform analyses using restricted data resources.

RESOURCE SHARING

Projects funded by Genome BC must also address sharing of resources generated by projects such as unique biological specimens and computer programs designed to analyze datasets. Biological reagents such as unique strains should be deposited into repositories such as ATCC and computer programs designed to analyze large datasets should be made available to others through the use of license agreements that adhere to "open source" principles (see for example, http://www.opensource.org/).

DATA DOCUMENTATION

Regardless of the mechanism used to share data, each dataset will require documentation. (Some fields refer to data documentation by other terms, such as metadata or codebooks.) Proper documentation is needed to ensure that others can use the dataset and to prevent misuse, misinterpretation, and confusion. Documentation provides information about the methodology and procedures used to collect the data, details about codes, definitions of variables, variable field locations, frequencies, and the like. The precise content of documentation will vary by scientific area, study design, the type of data collected, and characteristics of the dataset.

It is appropriate for scientific authors to **acknowledge the source of data** upon which their manuscript is based. Many investigators include this information in the methods and/or reference sections of their manuscripts. Journals generally include an acknowledgement section, in which the authors can recognize people who helped them gain access to the data. Authors using shared data should check the policies of the journal to which they plan to submit to determine the precise location in the manuscript for such acknowledgement.

INTERNATIONAL DATA REPOSITORIES

The following table provides examples of databases where various data types or unique resources produced by Genome BC-funded projects may be deposited. *Note that the following are intended as examples and this is not presented as an exhaustive, definitive list.*

Data type/Resource	Database
DNA, RNA, Protein, EST, STS, HTG Sequences	DDBJ/EMBL/GenBank
Gene expression data	Gene Expression Omnibus (GEO) SAGEmap
SNPs	dbSNP
Proteomics[1]	Not yet established (Researcher's website)
Protein structures	PDB
Interactions	IntAct/BIND
Software code	SourceForge.net/ http://gchelpdesk.ualberta.ca/
Biological strains	ATCC
Metabolomics	Not yet established (Researcher's website)
Publications	PubMed Central

1. Standards in proteomics are fluid but authors should adopt best practices such as the ontology based mzXML file format schema (http://tools.proteomecenter.org). When standardized databases emerge, researchers must deposit data into these databases.

OPEN ACCESS PUBLICATIONS

Final manuscripts are an important record of the research funded by Genome BC and open access to these publications is paramount. In order to foster open access to journal articles, Genome BC expects funded researchers to deposit a digital copy of their published manuscript and any appropriate supplementary information into PubMed Central (PMC) and to also provide Genome BC with a digital copy. Six months after the study's publication (or sooner if the publisher agrees) the manuscript will be made freely available to the public through PMC. If the publisher requests, the author's final version of the publication will be replaced in the PMC archive by the final publisher's copy with an appropriate link to the publisher's electronic database.

FUNDS FOR DATA SHARING

Genome BC recognizes that it takes time and money to prepare data for sharing. Thus, applicants can request funds for data sharing and archiving in their project proposal. Investigators who incorporate data sharing in the initial design of the study may more readily and economically establish adequate procedures for protecting the identities of participants and share a useful dataset with appropriate documentation.

DEFINITIONS

Restricted Data - datasets that cannot be distributed to the general public, because of, for example, participant confidentiality concerns, third-party licensing or use agreements, or national security considerations.

Data – The dataset and its associated documentation.

Data Archive – A place where machine-readable data are acquired, manipulated, documented, and distributed to others for further analysis and consumption.

Data Enclave – A controlled, secure environment in which eligible researchers can perform analyses using restricted data resources.

Data Federation – A group of data providers working under a common charter to serve data under uniform rules that govern data access and data use.

Data Sharing Agreement – A contract between data provider and data consumers that defines terms of data access and use. These terms should be consistently applied.

Final Research Data – Final research data are recorded factual materials commonly accepted in the scientific community and require documentation and validation. This data includes not only summary statistics or tables, but also all the data on which those statistics and tables are based. For the purposes of these guidelines, final research data do not include laboratory notebooks; partial datasets; preliminary analyses; drafts of scientific papers; plans for future research; peer review reports; communications with colleagues; or physical objects such as gels.

Metadata – The information that describes the data source and the time, place, and conditions under which the data were created. Metadata informs the consumer of who, when, what, where, why, and how data were generated. Metadata allows the data to be traced to a known origin and know quality.

Timeliness – In general, Genome BC considers the timely release and sharing of data to be no later than the acceptance for publication of the main findings from the final dataset. The Principal investigator will notify Genome BC of any delays to the release of data beyond the publication date or other scheduled dates for the release of data developed with Genome BC funds, such as the need to file provisional or full patent applications or apply for copy-write or trademark to protect intellectual property and/or know-how arising from the dataset.

Unique Data – Data that cannot be readily replicated. Examples of studies producing unique data include: large surveys that are too expensive to replicate; studies of unique populations, such as native or un-contacted peoples; studies conducted at unique times, such as after a natural disaster; studies conducted over long time scales; and studies of rare phenomena.