National Institutes of Health
Guidance for Investigators in Developing
Genomic Data Sharing Plans

The National Institutes of Health (NIH) Genomic Data Sharing (GDS) Policy[1] expects all investigators seeking NIH funding to develop a genomic data sharing plan if they are proposing research that will generate large-scale human and non-human genomic data.[2] This guidance document describes the type of information that should be provided in a genomic data sharing plan and when the plan should be submitted. The appendix to this document provides examples of genomic data sharing plans for human and non-human genomic research.[3]

Elements of a Genomic Data Sharing Plan

For extramural investigators, genomic data sharing plans are to be submitted as part of an application for funding. For all applicants proposing to generate human or non-human data, elements 1 and 2, a description of the data type and the data repository, should be provided at the time of the application. Applicants proposing to generate human data should also provide information addressing elements 3-5 and, if applicable, element 6 prior to award. Applicants proposing to generate non-human data need also to address element 3 prior to award. NIH intramural investigators should submit all relevant elements of the genomic data sharing plan to their NIH Institutes and Center (IC) Scientific Director (SD), or delegate, for review and approval. NIH intramural investigators should have in place an approved data sharing plan prior to start of the research.

1. **Data Type:** Explain whether the research being considered for funding involves human data, non-human data, or both. Denote the type of genomic data that will be shared (e.g., sequence, transcriptomic, epigenomic, and/or gene expression data) and whether it is individual-level data, aggregate-level data, or both. Also list any other information that is anticipated to be shared such as relevant associated data (e.g., phenotype or exposure data) and information necessary to interpret the data (e.g., study protocols, data collection instruments, survey tools).

2. **Data Repository:** Identify the data repositories to which the data will be submitted, and for human data, whether the data will be available through unrestricted[4] or controlled-access.[5] For human genomic data, investigators are expected to register all studies in the database of Genotypes and Phenotypes (dbGaP)[6] by the time data cleaning and quality control measures begin in addition to submitting the data to the relevant NIH-designated data repository (e.g., dbGaP, Gene Expression Omnibus (GEO), Sequence Read Archive (SRA), the Cancer Genomics Hub) after registration.

---

[1] NIH GDS Policy. See http://gds.nih.gov/PDF/NIH_GDS_Policy.pdf.

[2] The *Supplemental Information to the NIH Genomic Data Sharing Policy* provides examples of research generating large-scale genomic data subject to the Policy. See http://gds.nih.gov/pdf/supplemental_info_GDS_Policy.pdf.

[3] Note that if the proposed research falls within the scope of other NIH policies, such as the *NIH Policy on Sharing of Model Organisms for Biomedical Research* (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-04-042.html), investigators will be expected to fulfill the expectations of those policies.

[4] Data in unrestricted-access repositories (e.g., The 1000 Genomes Project) are publicly available to anyone.

[5] Controlled-access data (e.g., data in dbGaP) are made available for secondary research only after investigators have obtained appropriate approval to use the requested data for their proposed project.

[6] dbGaP. See http://www.ncbi.nlm.nih.gov/gap.

Non-human data may be made available through any widely used data repository, whether NIH-funded or not, such as GEO, SRA, Trace Archive, Array Express, Mouse Genome Informatics, WormBase, the Zebrafish Model Organism Database, GenBank, European Nucleotide Archive, or DNA Data Bank of Japan.

3. **Data Submission and Release Timeline:** Provide a timeline for sharing data in a timely manner. The *Supplemental Information to the GDS Policy*[2] provides expectations for the timelines of data submission and release based on the level of data processing. In general, NIH will release human genomic data no later than six months after the data have been submitted to NIH-designated data repositories and cleaned, or at the time of acceptance of the first publication, whichever occurs first, without restrictions on publication or other dissemination of research findings.

   Investigators should make non-human genomic data publicly available no later than the date of initial publication. However, availability before publication may be expected for certain data, projects (e.g., data from projects with broad utility as a resource for the scientific community such as microbial population-based genomic studies), or by the funding NIH IC.

4. **IRB Assurance of the Genomic Data Sharing Plan:** State whether an Institutional Review Board (IRB) or analogous review body has reviewed the genomic data sharing aspects of your project, or provide a timeline for such review. IRB review of the investigator's proposal for data submission is an element of the Institutional Certification[7] which assures that the proposal for data submission and sharing is appropriate. Please keep in mind that an Institutional Certification is generally required for extramural investigators prior to NIH grant award along with other Just-in-Time information or finalization of a contract. For NIH intramural investigators, an Institutional Certification memorandum should be completed and sent from the SD, or delegate, to the IC Genomic Program Administrator (GPA) before research is begun, whenever possible.

5. **Appropriate Uses of the Data:** The appropriate use of the data should be described. Under the GDS Policy, data is expected to be shared for broad research purposes. If such use of the data is not appropriate, as expressed in informed consent documents of the research participants whose data are included in the dataset, any limitations on the data use should be described in the Institutional Certification. NIH provides standard language[8] to guide the development of data use limitations.

6. **Request for an Exception to Submission:** If submission of human data generated in the study would not be appropriate because the Institutional Certification[9] criteria cannot be met, the investigator should explain why in the genomic data sharing plan and describe an alternative mechanism for data sharing. If the funding IC grants an exception to submission, the research will be registered in dbGaP and the reason for the exception and the alternative sharing plan will be described. For NIH intramural studies, the NIH Deputy Director for Intramural Research will make the final decision on the exception request, after the IC has made its determination.

[7] Points to Consider for Institutions and Institutional Review Boards in Developing Institutional Certifications for Submitting Human Data under the Genomic Data Sharing Policy. See http://gds.nih.gov/pdf/PTC_for_IRBs_and_Institutions.pdf.
[8] See http://gds.nih.gov/pdf/standard_data_use_limitations.pdf
[9] See http://gds.nih.gov/Institutional_Certifications.html

APPENDIX
Examples of Genomic Data Sharing Plans

Example 1: Data from human specimens not yet collected will be shared through NIH-designated data repositories.

Data generated from 800 human samples will be shared through unrestricted-access NIH-designated data repositories; individuals who do not give consent for sharing data will be excluded from the study. Genomic data include individual- and aggregate-level data from whole exome sequencing and genome-wide expression arrays. The study will be registered in dbGaP and the following data and information will be shared through the Sequence Read Archive and Gene Expression Omnibus:

- Study documents (e.g., study protocol, manual of operations, questionnaire, and data abstraction forms)
- Individual-level sequence data produced as part of Specific Aim 1 (i.e., files for single nucleotide polymorphisms)
- Individual-level expression data included in the analyses under Specific Aim 2 (i.e., array data and intensity peaks)
- Associated phenotypic data

The sequence and expression data will be shared once the data have been cleaned and quality control procedures are completed, which is expected to be completed no more than two months after the data have been generated. Data will be generated in years 1 and 2 and submitted in years 2 and 3 of the proposed study. The draft consent form provides consent for the data to be used for future research purposes and to be shared broadly through unrestricted-access databases. The Institutional Certification signed by the Institutional Signing Official will be submitted prior to award, along with any other Just-in-Time information.

The IRB advised that the sequence data produced through this award may be shared through unrestricted-access NIH-designated data repositories, consistent with data sharing under the NIH GDS Policy. The IRB will review the protocol of this project and will assure, prior to funding, that:

A. The protocol for the collection of genomic and phenotypic data is consistent with 45 CFR Part 46;[10]

B. Data submission and subsequent data sharing for research purposes are consistent with the informed consent of study participants from whom the data were obtained;

C. Consideration was given to risks to individual participants and their families associated with data submitted to NIH-designated data repositories and subsequent sharing;

D. To the extent relevant and possible, consideration was given to risks to groups or populations associated with submitting data to NIH-designated data repositories and subsequent sharing; and

[10] See http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html.

E. The investigator's plan for de-identifying datasets is consistent with the standards outlined in the GDS Policy.

Example 2: Data are generated from human specimens collected before the effective date of the GDS Policy, and the data will be shared through NIH-designated data repositories.

Genomic data will be generated from specimens that were previously collected from 2,000 study participants. The genotype and relevant phenotype data for participants will be shared through dbGaP, a controlled-access database, once the genotyping data have been cleaned, which we expect to be completed no more than two months after genotyping is finished. Submission of individual-level genome-wide genotype data produced as part of Specific Aim 1 and individual-level phenotypic data related to mood disorders included in the analyses under Specific Aim 2 is anticipated in year 2 of the proposed study.

The consent for the collection of specimens did not directly address the broad sharing of participants' data but did denote their desire to advance science. After careful review, the IRB determined that data submission was not inconsistent with the terms outlined in the consent. The Institutional Certification, which will be provided prior to award along with any other Just-in-Time information, will include the following DUL: "Use of these data is limited to health/medical/biomedical purposes, which does not include the study of population origins or ancestry."

The Institutional Review Board (IRB) advised that the genotyping data generated from 2,000 specimens may be shared through NIH-designated data repositories, consistent with data sharing under the NIH GDS Policy. The IRB has reviewed the study protocol and assures that:

A. The protocol for the collection of genomic and phenotypic data is consistent with 45 CFR Part 46;[10]

B. Data submission and subsequent data sharing for research purposes are consistent with the informed consent of study participants from whom the data were obtained;

C. Consideration was given to risks to individual participants and their families associated with data submitted to NIH-designated data repositories and subsequent sharing;

D. To the extent relevant and possible, consideration was given to risks to groups or populations associated with submitting data to NIH-designated data repositories and subsequent sharing; and

E. The investigator's plan for de-identifying datasets is consistent with the standards outlined in the GDS Policy.

Example 3: Data are generated from human specimens collected before the effective date of the GDS Policy, and the data cannot be shared through NIH-designated data repositories.

Genomic data from more than 100 genes in the genome will be generated from specimens previously collected from 700 study participants from a small population in Africa. The consent form did not directly address the broad sharing of participants' data nor the risks associated with broad data sharing of these data. Because of the small population and the lack of information in the consent form, the IRB concluded that it is not appropriate to share these individual-level data collected from existing specimens through any NIH-designated repository and is requesting an exception to data

deposition be granted. Pursuing a re-consent process for these participants is not a viable option due to the time lapse between acquiring the samples and generating the data. As an alternative data sharing plan, the University has agreed to share aggregate-level data that will be submitted to dbGaP and to provide a mechanism to facilitate data sharing through direct collaborations with other investigators under appropriate IRB oversight. The aggregate-level data will include aggregated minor allele frequencies and associated p-values. Other investigators may contact the principal investigator if interested in collaborating on a project that requires use of the individual-level data. All future research participants will be asked to sign an amended consent form that is consistent with the expectation of broad data sharing.

Example 4: Data from non-human specimens will be shared through NIH-designated data repositories.

The University will share individual-level genotype data from 1,500 mice by depositing these data in Sequence Read Archive, which is an NIH-funded repository. In addition, the study protocol, manual of operations, and phenotype data will be submitted. The genotype data will be made publicly available no later than the date of initial publication, which we anticipate during year 3 of the proposed research.