



This document is scheduled to be published in the Federal Register on 09/20/2013 and available online at <http://federalregister.gov/a/2013-22941>, and on [FDsys.gov](http://FDsys.gov)

**[Billing Code 4140-01-P]**

**DEPARTMENT OF HEALTH AND HUMAN SERVICES**

**National Institutes of Health**

**Draft NIH Genomic Data Sharing Policy Request for Public Comments**

**SUMMARY:** The National Institutes of Health (NIH) is seeking public comments on the draft Genomic Data Sharing (GDS) Policy that promotes sharing, for research purposes, of large-scale human and nonhuman genomic<sup>1</sup> data generated from NIH-supported and NIH-conducted research.

**DATES:** To ensure that your comments will be considered, please submit your response to this Request for Comments no later than 60 days after publication of this notice.

**ADDRESSES:** Submit comments by any of the following methods:

- Online: <http://gds.nih.gov/survey.aspx>
- Fax: 301-496-9839
- Mail/Hand delivery/Courier (for paper, disk, or CD-ROM submissions) to: Genomic Data Sharing Policy Team, Office of Science Policy, National Institutes of Health, 6705 Rockledge Drive, Suite 750, Bethesda, MD 20892.

**FOR FURTHER INFORMATION CONTACT:** Genomic Data Sharing Policy Team, Office of Science Policy, National Institutes of Health, 6705 Rockledge Drive, Suite 750, Bethesda, MD 20892, 301-496-9838, [GDS@mail.nih.gov](mailto:GDS@mail.nih.gov).

## **SUPPLEMENTARY INFORMATION**

### **Background**

The NIH's mission is to seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability. The draft GDS Policy supports this mission by promoting the sharing of genomic research data, which maximizes the knowledge gained. Not only does data sharing allow data generated from one research study to be used to explore a wide range of additional research questions, it also enables data from multiple projects to be combined, amplifying the scientific value of data many times. Broad research use of the data enhances public benefit by helping to speed discoveries that increase the understanding of biological processes that affect human health and the development of better ways to diagnose, treat, and prevent disease.

The NIH has promoted data sharing for many years, and in 2003, the NIH issued a general policy for sharing research data.<sup>2,3</sup> In 2007, the NIH issued a more specific policy to promote sharing of data generated through genome wide association studies (GWAS),<sup>4,5</sup> which examine thousands of single nucleotide polymorphisms (SNPs) across the genome to identify genetic variants that contribute to human diseases, conditions, and traits. To facilitate the sharing of genomic and phenotypic data from GWAS, the NIH created the database of Genotypes and Phenotypes (dbGaP) with a two-tiered system for distributing the data: open access, for data that are available to the public without restrictions, and controlled access for data that are made available only for research purposes that are consistent with the original informed consent under which the data were collected.

Not long after the GWAS policy was issued, advances in DNA sequencing and other high-throughput technologies, and a steep drop in DNA sequencing costs, enabled the NIH to fund research that generated even greater volumes of GWAS and other types of genomic data. In 2009, the NIH announced<sup>6</sup> its intention to extend the GWAS Policy to encompass data from a wider range of genomic research.

The draft GDS Policy applies to research involving nonhuman genomic data as well as human data that are generated through array-based and high-throughput genomic technologies (e.g., SNP, whole-genome, transcriptomic, epigenomic, and gene expression data). (See section II of the draft Policy.) The NIH considers access to such data particularly important because of the opportunities to accelerate research through the power of combining such large and information-rich datasets. The draft GDS Policy is aligned with Administration priorities and a recent directive to agencies to increase access to digital scientific data resulting from federally funded research.<sup>7</sup>

### **Overview of the Policy**

The draft GDS Policy describes the responsibilities of investigators and institutions for the submission of nonhuman and human genomic data to the NIH (section IV) and the use of controlled-access data (section V). The Policy also provides expectations regarding intellectual property (section VI).

When data sharing involves human data, the protection of research participant privacy and confidentiality is paramount, and the Policy reflects the NIH's continued commitment to responsible data stewardship, which is essential to uphold the public trust in biomedical research. The draft GDS Policy, like the GWAS Policy, includes a number of provisions to protect research participant privacy (see section IV.C). For example, prior to data submission, traditional identifiers such as name, date of birth, street address, and social security number should be removed. The de-identified<sup>8</sup> data are coded using a random, unique code to protect participant privacy. The NIH also maintains the expectation established under the GWAS Policy that the responsible Institutional Signing Official<sup>9</sup> of the submitting institution should provide an Institutional Certification to the funding NIH Institute or Center prior to award. An Institutional Certification assures that the data have been or will be collected in a legal and ethically

appropriate manner and have been de-identified. The draft GDS Policy clarifies the provisions of the Institutional Certification for datasets submitted to NIH-designated data repositories in Section IV.C.5.

The NIH expects the Policy to be effective 60 days after the publication of the final Policy.

### **Request for Comments**

As part of the process of developing the GDS Policy, the NIH encourages the public to provide comments on any aspect of the draft GDS Policy.

Comments should be submitted electronically to <http://gds.nih.gov/survey.aspx>.

Comments may also be submitted by fax (301-496-9839), or mailed to the Genomic Data Sharing Policy Team, Office of Science Policy, National Institutes of Health, 6705 Rockledge Drive, Suite 750, Bethesda, MD 20892.

Responding to this request for comments is voluntary. Submitted comments are considered public information; do not include any information that you wish to remain private and confidential. Comments in their entirety will be posted along with the submitter's name and affiliation on the NIH GDS website after the public comment period closes. Commenters will receive a confirmation acknowledging receipt of comments but will not receive individual feedback on any suggestions. Please note that the government will not pay for the use of any information contained in the response.

The NIH intends to hold one or more public webinars on the draft Policy. Information about the webinars will be made available at <http://gds.nih.gov>.

## **DRAFT NIH GENOMIC DATA SHARING POLICY**

### **I. Purpose**

The draft Genomic Data Sharing (GDS) Policy sets forth expectations that ensure the broad and responsible sharing of genomic research data. Sharing research data supports the NIH mission<sup>10</sup> and is essential to facilitate the translation of research results into knowledge, products, and procedures that improve human health. The NIH has longstanding policies to make data publicly available in a timely manner from the research activities that it funds.<sup>11,12</sup>

## **II. Scope and Applicability**

This Policy applies to all NIH-funded research that involves large-scale human and nonhuman genomic data produced by array-based or high-throughput genomic technologies, such as GWAS<sup>13</sup> SNP, whole-genome, transcriptomic, epigenomic, and gene expression data, irrespective of funding level and funding mechanism (i.e., grant, contract, or intramural support). Appendix A provides examples of research that are subject to the Policy. At appropriate intervals, the NIH will review the types of research to which this Policy may be applicable, and changes to the scope will be defined in supplementary materials to the final GDS Policy. Notification of any changes will be provided to investigators and institutions through standard NIH communication channels (*e.g., NIH Guide for Grants and Contracts*).

Compliance with this Policy will become a special term and condition in the Notice of Award or the Contract Award. Failure to comply with the terms and conditions of the funding agreement could lead to enforcement actions, including the withholding of funding, consistent with 45 CFR 74.62 and/or other authorities, as appropriate.

## **III. Effective Date**

The effective date of this Policy is [To Be Determined], and pertains to the following funding mechanisms:

- Competing grant applications<sup>14</sup> that are submitted to the NIH as of the [TBD] receipt date;

- Proposals for contracts that are submitted to the NIH as of [TBD]; and
- NIH intramural research projects that are approved as of [TBD].

#### **IV. Responsibilities of Investigators Submitting Genomic Data**

##### **A. Data Sharing Plans**

Investigators seeking NIH funding should contact appropriate Institute or Center (IC) Program or Project Officials<sup>15</sup> as early as possible to discuss data sharing expectations and timelines that would apply to their proposed studies. Investigators and their institutions are expected to address plans for following this Policy in the data sharing section of funding applications and proposals. Any resources needed to support a proposed data sharing plan should be included in the project's budget. NIH intramural investigators are expected to address data sharing plans with their IC scientific leadership prior to initiating applicable research and are encouraged to contact their IC leadership or the Office of Intramural Research for guidance.

##### **B. Nonhuman and Model Organism Genomic Data**

###### **1. Data Submission Expectations and Timeline**

Nonhuman data (including microbial and microbiome data) and data from large-scale genomic projects for model organisms<sup>16</sup> are to be shared in a timely manner. Investigators should make nonhuman and model organism data publicly available no later than the date of initial publication. However, certain data types or NIH research initiatives may expect an earlier data release (*e.g.*, microbial or microbiome data, or projects with broad utility as a resource for the scientific community). (See Appendix A for specific expectations for data submission and release.)

###### **2. Data Repositories**

Data should be made available through any widely used data repository, whether NIH-funded or not, such as the Gene Expression Omnibus (GEO),<sup>17</sup> Sequence Read Archive

(SRA),<sup>18</sup> Trace Archive,<sup>19</sup> Array Express,<sup>20</sup> Mouse Genome Informatics (MGI),<sup>21</sup> WormBase,<sup>22</sup> the Zebrafish Model Organism Database (ZFIN),<sup>23</sup> GenBank,<sup>24</sup> European Nucleotide Archive (ENA),<sup>25</sup> or DNA Data Bank of Japan (DDBJ).<sup>26</sup>

### **C. Human Genomic Data**

#### **1. Data Submission Expectations and Timeline**

Guidance to govern human genomic data submission timelines and data release expectations is provided in Appendix A. The NIH will release data submitted to NIH-designated data repositories without restrictions on publication or other dissemination no later than six months after the initial data submission to an NIH-designated data repository,<sup>27</sup> or at the time of acceptance of the first publication, whichever occurs first.

Human data that are submitted to NIH-designated data repositories should be de-identified according to the standards set forth in the HHS Regulations for the Protection of Human Subjects<sup>28</sup> and the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.<sup>29</sup> The de-identified data should be assigned a random, unique code, and the key held by the submitting institution.

The NIH encourages researchers and institutions submitting large-scale genomic datasets to NIH-designated data repositories to consider whether a Certificate of Confidentiality could serve as an additional safeguard to prevent compelled disclosure of any personally identifiable information that it may hold.<sup>30</sup> The NIH has obtained a Certificate of Confidentiality for dbGaP.<sup>31</sup>

#### **2. Data Repositories**

Applicable studies with human genomic data should be registered in the database of Genotypes and Phenotypes (dbGaP)<sup>32</sup> no later than the time that data cleaning and quality control measures begin. Investigators should submit human data to the relevant NIH-designated data repository (*e.g.*, dbGaP, GEO, SRA, the Cancer Genomics Hub<sup>33</sup>). NIH-designated data

repositories need not be the exclusive source for facilitating the sharing of genomic data.

Investigators who elect to submit data to a non-NIH-designated data repository should confirm that appropriate data security, confidentiality, and privacy measures are in place.

### 3. Tiered System for the Distribution of Human Data

Respect for and protection of the interests of research participants is fundamental to the NIH's stewardship of human genomic data. The informed consent under which the data or sample were collected is the basis for the submitting institution to determine the appropriateness of data submission to NIH-designated data repositories, and whether the data should be available through open or controlled access. Controlled-access data in NIH-designated data repositories are made available for secondary research only after investigators have obtained approval from the NIH to use the requested data for a particular project. Open-access data are publicly available without restriction (e.g., The 1000 Genomes Project<sup>34</sup>).

### 4. Informed Consent

Submitting institutions, through their Institutional Review Boards (IRBs), are to review the informed consent materials for studies that are to be submitted to NIH-designated data repositories to determine whether the data are appropriate for sharing for secondary research use. Specific considerations may vary with the type of study and whether the data are obtained through prospective or retrospective data collections. The NIH provides additional information on issues related to the respect for research participant interests in its *Points to Consider for IRBs and Institutions in their Review of Data Submission Plans for Institutional Certifications*.<sup>35</sup> This and other policy-related documents will be updated once the Policy is final.

For studies initiated after the effective date of this Policy, the NIH expects the informed consent process and documents to state that a participant's genomic and phenotypic data may be shared broadly for future research purposes and also explain whether the data will be shared through open or controlled access. If human genomic data are to be shared in open-access



repositories, the NIH expects that participants will have provided explicit consent for sharing their data through open-access mechanisms. For studies proposing to use cell lines or clinical specimens,<sup>36</sup> the NIH expects that informed consent for future research use and broad data sharing will have been obtained even if the cell lines or clinical specimens are de-identified. If there are compelling scientific reasons that necessitate the use of cell lines or clinical specimens that were created or collected after the effective date of this Policy and that lack consent for research use and data sharing, investigators should provide a justification for the use of any such materials in the funding request.

For studies using data or specimens collected before the effective date of this Policy, there may be considerable variation in the extent to which data sharing and future genomic research was addressed within the informed consent materials for the primary research. In these cases, an assessment by an IRB, Privacy Board, or equivalent group is essential to ensure that data submission is not inconsistent with the informed consent provided by the research participant.

The NIH will accept data derived from cell lines or clinical specimens lacking consent for research use that were created or collected before the effective date of this Policy. Grandfathered genomic data that are currently available through open access may be submitted to an open-access NIH-designated data repository; otherwise, the data should be submitted to a controlled-access NIH-designated data repository.

While the NIH encourages broad access to genomic data, in some circumstances broad sharing may be inconsistent with the informed consent of the research participants whose data are included in the dataset. In such circumstances, institutions planning to submit aggregate- or individual-level data to the NIH for controlled access should note any data use limitations in the data sharing or data management plan submitted as part of the funding request. These data use limitations should be specified in the Institutional Certification submitted to the NIH prior to award.

## 5. Institutional Certification

The responsible Institutional Signing Official of the submitting institution should provide an Institutional Certification to the funding IC prior to award. The Institutional Certification should indicate whether the data will be submitted to an open- or controlled-access database and assure that:

- The data submission is consistent with applicable laws, regulations, and institutional policies;<sup>37</sup>
- The appropriate research uses of the data and any uses that are specifically excluded in the informed consent documents are delineated<sup>38</sup>
- The identities of research participants will not be disclosed to NIH-designated data repositories; and
- An IRB, Privacy Board, and/or equivalent body<sup>39</sup> has reviewed the investigator's proposal for data submission and assures that:
  - The protocol for the collection of genomic and phenotypic data was consistent with 45 CFR Part 46.
  - Data submission and subsequent data sharing for research purposes are consistent with the informed consent of study participants from whom the data were obtained;<sup>40</sup>
  - Risks to individuals and their families associated with data submitted to NIH-designated data repositories were considered;
  - To the extent relevant and possible, risks to groups or populations associated with data submitted to NIH-designated data repositories were considered; and
  - The investigator's plan for de-identifying datasets is consistent with the standards outlined in this Policy (see section IV.C.1.).

Institutions should indicate in the certification whether aggregate genomic data from datasets with data use limitations may be appropriate for general research use (i.e., use for any

research question such as research to understand the biological mechanisms underlying disease, development of statistical research methods, the study of populations origins). If so, the aggregate genomic data will be made available through the controlled-access compilation of aggregate genomic data<sup>41</sup> to facilitate secondary research.

#### 6. Data Withdrawal

Submitting investigators and their institutions may request removal of data on individual participants from NIH-designated data repositories in the event that a research participant withdraws his or her consent. However, data that have been distributed for approved research use cannot be retrieved.

#### 7. Exceptions to Data Submission Expectations

The NIH acknowledges that in some cases, circumstances beyond the control of investigators may preclude submission of data to NIH-designated data repositories (*e.g.*, country or state laws that prohibit data submission to a U.S. federal database). In such cases, investigators should provide a justification for any exceptions requested in the application or proposal. The funding IC may grant an exception to the submission of relevant data to the NIH, and the investigator would be expected to develop a plan to share data through other mechanisms. For transparency purposes, when exceptions are granted, studies will still be registered in dbGaP and the reason for the exception will be included in the registration record. Information about current expectations for exception requests will be made available on the GDS website.

### **V. Responsibilities of Investigators Accessing and Using Genomic Data**

#### **A. Requests for Controlled-Access Data**

Access to human data is through a two-tiered model involving open- and controlled-data access mechanisms. Requests for controlled-access data<sup>42</sup> are reviewed by NIH Data Access Committees (DACs).<sup>43</sup> DAC decisions are based primarily upon conformance of the proposed research as described in the access request to the data use limitations established by the

submitting institution through the Institutional Certification. The NIH DACs will accept requests for proposed research uses beginning one month prior to the anticipated data release date. The access period for all controlled-access data is one year; at the end of each approved period, data users can request an additional year of access or close out the project.

Investigators approved to download controlled-access data from NIH-designated data repositories and their institutions are expected to abide by the NIH User Code of Conduct<sup>44</sup> through their agreement to the Data Use Certification.<sup>45</sup> The Data Use Certification, co-signed by the investigators requesting the data and their Institutional Signing Official, specifies the terms and conditions for the secondary research use of controlled-access data, such as:

- Using the data only for the approved research;
- Protecting data confidentiality;
- Following all applicable laws, regulations, and local institutional policies and procedures for handling genomic data;
- Not attempting to identify individual participants from whom the data were obtained;
- Not selling any of the data obtained from the NIH-designated data repositories;
- Not sharing any of the data obtained from the NIH-designated data repositories with individuals other than those listed in the data access request;
- Agreeing to the listing of a summary of approved research uses in dbGaP along with the investigator's name and organizational affiliation;
- Agreeing to report, in real time, violations of the GDS Policy to the appropriate DAC;
- Providing annual updates on research using controlled-access datasets.

For investigators who are approved to use the data, the NIH maintains guidance on security practices<sup>46</sup> that outlines expected data security protections (e.g., physical security measures and user training) to ensure that the data are kept secure and not released to any person not permitted to access the data.

## **B. Acknowledgment Responsibilities**

The NIH expects all investigators who access genomic datasets from NIH-designated data repositories to acknowledge in all resulting oral or written presentations, disclosures, or publications the contributing investigator(s) who conducted the original study, the funding organization(s) that supported the work, the specific dataset(s) and applicable accession number(s), and the NIH-designated data repositories through which the investigator accessed any data.

## **VI. Intellectual Property**

Naturally occurring DNA sequences are not patentable in the United States.<sup>47</sup> Therefore, basic sequence data and certain related information (e.g., genotypes, haplotypes, p values, allele frequencies) are pre-competitive, and such data made available through NIH-designated data repositories and all conclusions derived directly from them should remain freely available, without any licensing requirements, for uses such as markers for developing assays and guides for identifying new potential targets for drugs, therapeutics, and diagnostics. In addition, the NIH discourages the use of patents to prevent the use of or block access to genomic or genotype-phenotype data developed with NIH support. The NIH encourages broad use of NIH-funded genomic data that is consistent with a responsible approach to management of intellectual property derived from downstream discoveries, as outlined in the NIH *Best Practices for the Licensing of Genomic Inventions*<sup>48</sup> and Research Tools Policy.<sup>49</sup> The NIH encourages patenting of technology suitable for subsequent private investment that may lead to the development of products that address public needs.

## **APPENDIX A**

### **Supplemental Information for the NIH Genomic Data Sharing Policy**

## **Overview**

This document provides additional guidance on the types of research projects to which the Genomic Data Sharing (GDS) Policy applies and the NIH's expectations for data submission and release.

## **Examples of Types of Research Covered Under the GDS Policy**

The GDS Policy is applicable to any NIH-funded research project involving nonhuman organisms or human specimens that produces genomic, metagenomic, epigenomic, or transcriptomic data from large-output sequencing instruments or genotyping platforms, such as projects that involve:

- Sequence data from tens of isolates from infectious organisms.
- Sequencing more than one gene or gene-sized region in more than 100 participants.
- More than 10,000 genes or regions from one participant (*e.g.*, whole genome sequencing).
- More than 100,000 variant sites in more than 100 participants.

## **Expectations for Data Submission and Data Release**

Data submitted to NIH-designated data repositories undergo different levels of data processing, and the expectations for data submission and data release are based on those levels. The table and text below describe the expectations for each level. The NIH will review these expectations at regular intervals, and any updates will be published on the GDS website and the research community will be notified through appropriate communication methods (*e.g.*, *The NIH Guide for Grants and Contracts*).

<b>Level</b>	<b>General Description of Data Processing</b>	<b>Example Data Types</b>	<b>Data Submission Expectation</b>	<b>Data Release Timeline</b>
0	Raw data generated	Instrument image	Not expected	NA

Level	General Description of Data Processing	Example Data Types	Data Submission Expectation	Data Release Timeline
	directly from the instrument platform	data		
1	Initial sequence reads, the most fundamental form of the data after the basic translation of raw input	DNA sequencing reads, ChIP-Seq reads, RNA-Seq reads, SNP arrays, arrayCGH	Not expected for human data if reads are included in Level 2 aligned sequence file (e.g., BAM)  Nonhuman de novo sequence data	NA  Up to 6 months for nonhuman data
2	Data after an initial round of analysis or computation to clean the data and assess basic quality measures	DNA sequence alignments to a reference sequence or de novo assembly, RNA expression profiling	Project specific, generally within 3 months after data generation	Up to 6 months after data submission or at the time of acceptance of the first publication, whichever occurs first
3	Analysis to identify genetic variants, gene expression patterns, or other features of the dataset	SNP or structural variant calls, expression peaks, epigenomic features	Project specific, generally within 3 months after data generation	Up to 6 months after data submission or at the time of acceptance of the first publication, whichever occurs first
4	Final analysis that relates the genomic data to phenotype or other biological states	Genotype-phenotype relationships, relationships of RNA expression or epigenomic patterns to biological state	Data submitted as analyses are completed	Data released with publication

Level 0 and level 1 data are the raw images and initial sequence reads, respectively, and have limited value to secondary data users. NIH policy does not expect submission of these data. An exception is made for de novo sequencing of nonhuman organisms unless those read data are provided within the level 2 submission. In the case of de novo sequencing for nonhuman organisms, investigators who are submitting level 1 data may request a holding period, not to exceed six months, during which the datasets will not be released for use by other investigators. For data submitted to NIH-designated data repositories, provisions may be made for creating an

exchange area in which such datasets may be shared among investigative teams prior to general release.

Submission of array-based data, such as gene expression, ChIP-chip, ArrayCGH, and SNP arrays can be submitted to GEO as level 1 data, which will not be accessible until a manuscript describing the data is published. It is the submitter's responsibility to ensure that the data and files submitted to GEO protect participant privacy in accordance with all applicable laws, regulations, and institutional policies, including the GDS Policy.

Level 2 constitutes a computational analysis in the form of higher order assembly or placement of the sequencing reads on a reference template. For human sequencing projects, the level 2 file comprises the reads "piled" on a reference human genome. A submission would be a file (*e.g.*, binary alignment matrix (BAM) files) usually containing the unmapped reads as well. GWAS and other types of projects (*e.g.*, RNA expression profiling or de novo sequencing) would also generate a level 2 placement or assembly file.

Generation of data files at level 2 generally requires substantial analysis and quality checks relating to both breadth of coverage of the targeted region and accuracy of assembly. Sufficient time will be allowed to complete the analysis and generate the assembly, up to the coverage and quality thresholds specified by a project or investigative team. In general, it is anticipated that this work could reasonably be completed within three months, and data submission would follow shortly thereafter. Data files may be held in an exchange area accessible only to the submitting investigators and collaborators for a period not to exceed six months from the time of submission. Following this period of exclusivity, the data will be available for research access without restrictions on publication.

Phenotype or clinical data should be submitted to the NIH-designated data repository at the earliest opportunity, but no later than the date of level 2 genomic data submission (or levels 2 and 3 for GWAS datasets), especially for studies in which all phenotype data have already been gathered. For studies in which phenotype data collections are ongoing and/or may be regularly



updated, data files should be submitted to NIH-designated data repositories as early as possible considering the practical needs for ensuring data accuracy; generally speaking, this time should not exceed six months after data collection.

Level 3 includes analysis to identify variants or to elucidate other features of the genomic dataset, such as gene expression patterns in an RNAseq assay. Level 3 data may be generated from a single level 2 data file (*e.g.*, variant sites versus the human reference genome), but will often derive from a compilation of sequencing assemblies (*e.g.*, in a genome study of a specific cancer type). Data submission expectations for level 3 files will vary substantially by project and therefore will require consultation with NIH program staff. As in level 2 data submission, level 3 files will be date stamped and the data producer may request a period of exclusivity not to exceed six months, after which time the datasets will be released through open- or controlled-access mechanisms as appropriate and without publication limitations.

Level 4 constitutes the final analysis, relating the genomic datasets to phenotype or other biological states as pertinent to the research objective. Data in this level are the project findings or the publication dataset. Investigators should submit these data prior to publication, and the data will be released concurrent with publication.

[FR Doc. 2013-22941 Filed 09/19/2013 at 8:45 am; Publication Date: 09/20/2013]

---

## References

<sup>1</sup> The genome is the entire set of genetic instructions found in a cell. See <http://ghr.nlm.nih.gov/glossary=genome>.

<sup>2</sup> Final NIH Statement on Sharing Research Data. February 26, 2003. See <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>.

<sup>3</sup> NIH Intramural Policy on Large Database Sharing. April 5, 2002. See <http://sourcebook.od.nih.gov/ethic-conduct/large-db-sharing.htm>.

<sup>4</sup> Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS). August 28, 2007. See <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html>.

<sup>5</sup> A GWAS is defined as any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition.

<sup>6</sup> Notice on Development of Data Sharing Policy for Sequence and Related Genomic Data. October 19, 2009. See <http://grants.nih.gov/grants/guide/notice-files/NOT-HG-10-006.html>.

<sup>7</sup> Office of Science and Technology Policy Memorandum, Expanding Public Access to the Results of Federally Funded Research. February 22, 2013. See <http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research>.

<sup>8</sup> “De-identified” refers to removing information that could be used to associate a dataset or record with a human individual. Under this Policy, data should be de-identified according to the standards set forth in the HHS Regulations for the Protection of Human Subjects and the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. The HIPAA Privacy Rule lists 18 identifiers that must be removed to classify data as de-identified. For the full list, see [http://privacyruleandresearch.nih.gov/pr\\_08.asp](http://privacyruleandresearch.nih.gov/pr_08.asp).

<sup>9</sup> An Institutional Signing Official is generally a senior official at an institution who is credentialed through the NIH eRA Commons system and is authorized to enter the institution into a legally binding contract and sign on behalf of an investigator who has submitted data or a data access request to the NIH.

<sup>10</sup> The NIH’s mission is to seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability. See <http://www.nih.gov/about/mission.htm>.

<sup>11</sup> Final NIH Statement on Sharing Research Data. February 26, 2003. See <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>.

<sup>12</sup> NIH Intramural Policy on Large Database Sharing. April 5, 2002. See <http://sourcebook.od.nih.gov/ethic-conduct/large-db-sharing.htm>.

<sup>13</sup> GWAS has the same definition in this policy as in the 2007 GWAS Policy: a study in which the density of genetic markers and the extent of linkage disequilibrium should be sufficient to capture (by the  $r^2$  parameter) a large proportion of the common variation in the genome of the population under study, and the number of samples (in a case-control or trio design) should provide sufficient power to detect variants of modest effect.

<sup>14</sup> Competing grant applications encompass all activities with a research component, including but not limited to the following: Research Grants (Rs), Program Projects (Ps), Cooperative Research Mechanisms (Us), Career Development Awards (Ks), and SCORs and other S grants with a research component.

<sup>15</sup> Investigators should refer to funding announcements or IC websites for contact information.

<sup>16</sup> NIH Policy on Sharing of Model Organisms for Biomedical Research. Release Date May 7, 2004. See <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-04-042.html>.

<sup>17</sup> Gene Expression Omnibus at <http://www.ncbi.nlm.nih.gov/geo/>.

<sup>18</sup> Sequence Read Archive at <http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?>

<sup>19</sup> Trace Archive at <http://www.ncbi.nlm.nih.gov/Traces/trace.cgi>.

<sup>20</sup> Array Express at <http://www.ebi.ac.uk/arrayexpress/>.

<sup>21</sup> Mouse Genome Informatics at <http://www.informatics.jax.org/>.

<sup>22</sup> WormBase at <http://www.wormbase.org>.

<sup>23</sup> The Zebrafish Model Organism Database at <http://zfin.org/>.

<sup>24</sup> GenBank at <http://www.ncbi.nlm.nih.gov/genbank/>.

<sup>25</sup> European Nucleotide Archive at <http://www.ebi.ac.uk/ena/>.

<sup>26</sup> DNA Data Bank of Japan at <http://www.ddbj.nig.ac.jp/>.

<sup>27</sup> A period for data preparation is anticipated prior to data submission to the NIH, and the appropriate time intervals for that data preparation (or data cleaning) will be subject to the particular data type and project plans (see Appendix A). Investigators should work with NIH Program or Project Officials for specific guidance.

<sup>28</sup> See 45 CFR 46.102(f) at <http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.102>.

<sup>29</sup> See 45 CFR 164.514(b)(2). The list of HIPAA identifiers that must be removed is available at: <http://www.gpo.gov/fdsys/pkg/CFR-2002-title45-vol1/pdf/CFR-2002-title45-vol1-sec164-514.pdf>.

<sup>30</sup> For additional information about Certificates of Confidentiality, see <http://grants.nih.gov/grants/policy/coc/>.

- <sup>31</sup> Confidentiality Certificate. HG-2009-01. Issued to the National Center for Biotechnology Information, National Library of Medicine, NIH. See [http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document\\_name=ConfidentialityCertificate.pdf](http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document_name=ConfidentialityCertificate.pdf).
- <sup>32</sup> database of Genotypes and Phenotypes at <http://www.ncbi.nlm.nih.gov/gap>.
- <sup>33</sup> Cancer Genomics Hub at <https://cghub.ucsc.edu/>.
- <sup>34</sup> The 1000 Genomes Project at <http://www.1000genomes.org/>.
- <sup>35</sup> Points to Consider for IRBs and Institutions in their Review of Data Submission Plans for Institutional Certifications. See [http://gwas.nih.gov/pdf/PTC\\_for\\_IRBs\\_and\\_Institutions\\_revised5-31-11.pdf](http://gwas.nih.gov/pdf/PTC_for_IRBs_and_Institutions_revised5-31-11.pdf).
- <sup>36</sup> Clinical specimens are specimens that have been obtained through clinical practice.
- <sup>37</sup> For the submission of data derived from cell lines or clinical specimens lacking research consent that were created or collected before the effective date of this Policy, the Institutional Certification needs to address only this item.
- <sup>38</sup> For guidance on clearly communicating inappropriate data uses, see NIH Points to Consider in Drafting Effective Data Use Limitation Statements, [http://gwas.nih.gov/pdf/NIH\\_PTC\\_in\\_Drafting\\_DUL\\_Statements.pdf](http://gwas.nih.gov/pdf/NIH_PTC_in_Drafting_DUL_Statements.pdf).
- <sup>39</sup> “Equivalent body” is used here to acknowledge that some primary studies may be conducted abroad and in such cases the expectation is that an analogous review committee to an IRB or Privacy Board (*e.g.*, Research Ethics Committees) may be asked to participate in the presubmission review of proposed genomic projects.
- <sup>40</sup> As noted earlier, for studies using data or specimens collected before the effective date of this Policy, the IRB or Privacy Board should review informed consent materials to ensure that data submission is not inconsistent with the informed consent provided by the research participants.
- <sup>41</sup> Compilation of Aggregate Genomic Data. dbGaP study accession: phs000501.v1.p1. See [http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study501.cgi?study\\_id=phs000501.v1.p1&pha=&phaf=](http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study501.cgi?study_id=phs000501.v1.p1&pha=&phaf=).
- <sup>42</sup> dbGaP Authorized Access. See <https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>.
- <sup>43</sup> For a list of NIH Data Access Committees, see [http://gwas.nih.gov/04po2\\_1DAC.html](http://gwas.nih.gov/04po2_1DAC.html).
- <sup>44</sup> User Code of Conduct. See [https://dbgap.ncbi.nlm.nih.gov/aa/GWAS\\_Code\\_of\\_Conduct.html](https://dbgap.ncbi.nlm.nih.gov/aa/GWAS_Code_of_Conduct.html).
- <sup>45</sup> Model Data Use Certification Agreement. See [http://gwas.nih.gov/pdf/Model\\_DUC\\_7-26-13.pdf](http://gwas.nih.gov/pdf/Model_DUC_7-26-13.pdf).
- <sup>46</sup> Security Best Practices. See [http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document\\_name=dbgap\\_2b\\_security\\_procedures.pdf](http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document_name=dbgap_2b_security_procedures.pdf).
- <sup>47</sup> In Association for Molecular Pathology et al. v. Myriad Genetics, Inc., et al. 569 U.S. \_\_\_\_ 2013. See [http://www.supremecourt.gov/opinions/12pdf/12-398\\_1b7d.pdf](http://www.supremecourt.gov/opinions/12pdf/12-398_1b7d.pdf).
- <sup>48</sup> NIH Best Practices for the Licensing of Genomic Inventions. See [http://www.ott.nih.gov/policy/genomic\\_invention.html](http://www.ott.nih.gov/policy/genomic_invention.html).
- <sup>49</sup> Research Tools Policy. See [http://www.ott.nih.gov/policy/research\\_tool.aspx](http://www.ott.nih.gov/policy/research_tool.aspx).

Date: September 16, 2013.

---

Lawrence A. Tabak,

Deputy Director

National Institutes of Health