**National Cancer Institute (NCI)**
**Framework for Implementing the National Institutes of Health (NIH)**
**Genomic Data Sharing (GDS) Policy**
*November 13, 2015*

**INDEX**

## I. Purpose

- NCI supports and complies with all NIH data sharing policies. This document provides a framework to promote consistency across the Institute with regard to the NIH Genomic Data Sharing (GDS) Policy[1] implementation.
- NCI will update this implementation framework as needed to maintain consistency with NIH policies and implementation guidance.
- This guidance will be reviewed and updated on a regular basis to reflect technological changes and to balance the benefits to the community with the costs and level of effort required to share, store, curate, and provide access to the data.

***Overarching Principles***

- Broad data sharing promotes maximum public benefit from federally funded genomics research.
- NCI supports the broadest appropriate genomic data sharing with timely data release through broadly accessible open or, if more appropriate, controlled access data repositories [*e.g.*, the database of Genotypes and Phenotypes (dbGaP[2]), or the Genomic Data Commons].
- Systems to ensure robust participant protection and appropriate oversight of research conduct, data quality, data management, data sharing, and data use are fundamental to effective data sharing policies and practices.
- Data sharing allows 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.
- Data to be shared should be annotated to enable data reuse, understanding, harmonization and meta-analysis.

## II. Effective Dates

- The GDS Policy applies to competing grant applications and contract proposals submitted to NIH on or after January 25, 2015.
- The first round of grant applications submitted for the January 2015 cycle would likely receive funding in the late summer or early fall of 2015. To align with this timeline, the intramural programs will begin submitting data generated on or after August 31, 2015.

## III. Scope and Applicability

- The GDS Policy applies to all NIH-funded research that generates large-scale human or non-human genomic data as well as the use of these data for subsequent research.
- Large-scale data include data from genome-wide association studies (GWAS) and single nucleotide polymorphism (SNP) arrays, as well as genome sequence, transcriptomic, metagenomic, epigenomic, and gene expression data, irrespective of funding level and funding mechanism (e.g., grant, contract, cooperative agreement, or intramural support).[3]

---

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

[2] Database of Genotypes and Phenotypes (http://www.ncbi.nlm.nih.gov/gap)

[3] Studies of smaller populations (*e.g.*, family studies or rare disease studies) should follow this principle wherever feasible and appropriate based on an IRB review of the study design and consent processes.

- Examples of research projects involving genomic data that will be subject to the Policy can be found in Supplement 1. Examples of research outside the scope of the GDS Policy are also provided in Supplement 1.
- Data sharing priorities will be set by NCI leadership based on the state of the science, programmatic priorities, and resource availability.

## IV. Data Standards

### *Minimal information guidelines*

- Data being shared are expected to meet basic technical quality standards appropriate to the field of study.
- Metadata around the study and annotations that are necessary to reproduce any published table or analysis must be included with genomic data submissions. This includes all relevant study information, materials and methods, and analytic methods. Metadata and annotations should conform to the recommended standards, outlined below.
- The specimen acquisition and experimental procedures as well as the data processing and analysis methods (such as alignment algorithms, software versions, etc.) are required with data submission.
- Information or data pertinent to the mining of genomic data—such as associated phenotype data (*e.g.*, clinical information), exposure data, and descriptive information (*e.g.*, protocols or methodologies used)—should be shared.

### *Data repositories*

- To the extent possible, genomic data should be shared via a supported NIH repository. Acceptable repositories are listed on the NIH Genomic Data Sharing Data Repository list[4].
- Extramural funding announcements may indicate a specific data repository to be used.
- If an investigator needs to share data for which there is not an appropriate data repository (a new data type not supported, for example), the investigator should contact the appropriate NCI Genomic Program Administrator (listed in Supplement 2) as early as possible to discuss alternatives, ideally prior to data production.
- All studies, regardless of which repository is used, will be registered in dbGaP where a basic summary of the study and information on how to request access, will be listed.

### *Recommended use of data standards*

- Data reuse is facilitated when the data conform to accepted data standards because the steepness of the learning curve for researchers is reduced and potential errors from misunderstanding of the data or metadata are minimized and analytic pipelines can be re-used.
- For this reason, NCI strongly encourages depositors to GDS repositories to utilize existing, well-documented data standards.
- Terms for disease, cell type, and tissue type as well as other annotations should be linked to NCI Thesaurus (NCIt)[5] concept identifiers when they exist and to other identifiers such as a Uniform Medical Language System (UMLS)[6] Unique Identifier and a term from an existing ontology[7] when an NCIt identifier is not available.

[4] GDS Data Repositories (http://gds.nih.gov/02dr2.html)
[5] NCI Thesaurus (http://ncit.nci.nih.gov/)
[6] UMLS (http://www.nlm.nih.gov/research/umls/)
[7] Bioportal (http://bioportal.bioontology.org)

- Wherever possible, existing common data elements should be used[8]. For clinical specimens, the data elements that would be included in reporting to clinicaltrials.gov are required.
- **For specific guidance on expected data formats, please refer to Supplement 3.**

## V. Data Sharing Plans (DSP)

- In determining appropriate data sharing plans for NCI research programs or projects, the following elements will be considered by extramural program staff and intramural leadership:
- The potential value of the dataset for use in secondary analyses to confirm findings, explore different research questions, and develop or refine analytic methodologies or programs;
    - Costs and other resource issues pertaining to data deposition, management, or access needs (*e.g.*, the availability of appropriate public data repositories or other data sharing mechanisms).
- For studies involving human data, the additional elements below will be considered in the development of data sharing plans:
    - Research participant informed consent [including Institutional Review Board (IRB) assessments of informed consent processes and consent documents with regard to broad data sharing and future research use];
    - Participant protection concerns (e.g., participant privacy or potential for group harm).
- The NCI expects that DSPs (see example in Supplement 4) will be collected and reviewed at the earliest time possible. For example for extramural programs a DSP should be included at the time of the "Awaiting Receipt of Application" (ARA) submission.
- **Extramural program staff and intramural program leadership, as applicable, should monitor the progress of each project falling under the GDS policy throughout its life cycle to ensure that data sharing progress is consistent with the DSP and be proactive in taking action to address any issues that arise.**

***Extramural programs***

- Extramural researchers should include data sharing plans in their grant or contract applications **in the Resource Sharing section of the grant application** and report on data sharing progress (addressing key milestones) through annual progress reports as appropriate.
- **NCI Program Directors will assess whether the project falls within the scope of the GDS policy and, if so, whether the DSP is adequate. Program Directors must approve the DSP prior to funding.**
- Program staff will evaluate data sharing plans according to the principles and expectations defined in the NIH GDS Policy[9] and this Framework, in consultation with their NCI Genomic Program Administrator (GPA) (Supplement 2), as appropriate.
- The Program Director (PD) should discuss DSP requirements with potential applicants as early in the pre-award process as possible (e.g., at the time of ARA submission). The PD should send any questions to the division or center's GPA[10] for clarification (Supplement 2).
- Program staff may require that the applicant provide a more detailed DSP during the pre-award period. If needed, alternative data sharing plans may be negotiated with applicants before funding is provided; these plans may be factored into assessments regarding program priority for funding decisions.

[8] See https://cdebrowser.nci.nih.gov/CDEBrowser/ and http://www.nlm.nih.gov/cde/
[9] NIH Guidance for Investigators Developing Data Sharing Plans
http://gds.nih.gov/pdf/NIH_guidance_developing_GDS_plans.pdf
[10] GPA (http://gds.nih.gov/04po2_2GPA.html)

- If different from the DSP submitted in the Resource Sharing section of the grant application, the finalized DSP should be included with Just-in-Time (JIT) materials. All investigators generating data covered by GDS Policy have in place a DSP approved by the program director prior to funding.
- All approved data sharing plans should be forwarded to an NCI GPA for review to promote consistency across the Institute, identify areas in need of further discussion, and ensure that the DSP contains all relevant information for initial registration of the study in dbGaP.
- Program Directors will review progress updates for data sharing and work with investigators to ensure compliance.

***Intramural programs***

- The Scientific Director or an appointed delegate, and GPA should review DSP at the earliest time possible. The study organism (e.g. human or non-human) and how scientific review takes place within the NCI intramural research program will dictate when this will occur:
    - *Prospective scientific review* - The DSP should be approved by the SD (or delegate) and GPA before the funding decision is made.
    - *Retrospective scientific review (e.g., quadrennial site visits)* – The DSP should be approved by the SD (or delegate) prior to data generation.

## VI. Institutional Certification (IC)

- Institutions are responsible for assuring (through an Institutional Certification) that plans for the submission of human genomic data to the NIH meet the expectations of the GDS Policy.
- The data submitting institutions (including the NCI Intramural Research Program (IRP)) should submit an Institutional Certification[11] in accord with all terms outlined in the NIH GDS Policy. Specifically:
    - For studies that initiate participant recruitment after the implementation date of the NIH GDS Policy and this Framework, submitting institutions should assure that future research use and data sharing are consistent with the informed consent provided by study participants.
    - For studies where participant recruitment occurred prior to the effective date of the NIH GDS Policy, submitting institutions should assure that future research use and data sharing are not inconsistent with the informed consent provided by study participants.
    - If established or commercially available cell lines or tissue samples are to be used as data sources within the study, investigators should seek whenever possible such sources where consent for future research use and data sharing can be documented.
    - Phenotype or clinical variables submitted for data release may be adjusted to promote participant privacy or other participant protection concerns as assessed by an IRB.
    - For studies where participant recruitment occurred after the effective date of the NIH GDS Policy, study participants should be consented for broad data sharing.
- An Institutional Certification must accompany the submission of all human data to the NIH Database of Genotypes and Phenotypes (dbGaP). NOTE: Currently the IC templates[11] (formats) are specific to either an extramural institution or an intramural one.
- For multi-center studies with samples collected at several institutions, the NIH understands that the submitting institution is not necessarily the IRB of record for all sites. However, the submitting institution should assure the NIH that based on either its own review or assurance from other institutions, the expectations of the Policy are met for the entire dataset. Institutions may choose to collect and submit a single site IC memo from each site contributing samples or submit a multi site IC memo.

[11] Institutional Certifications (http://gds.nih.gov/Institutional_Certifications.html)

- For studies involving human data, the responsible Institutional Signing Official (SO)[12] of the submitting institution should provide an Institutional Certification to the NCI, ideally prior to award.
- To ensure the appropriateness of data sharing prior to a funding decision and to facilitate sharing after data have been generated, submission of the Institutional Certification memo(s) is expected at the earliest time possible.

***Extramural programs***

- IC memo(s) should be completed and signed by the study Principal Investigator (PI) and an authorized Institutional SO and included in the grant or contract application, submitted with Just-in-Time (JIT) materials, or at the latest included with the year-1 Research Performance Progress Report (RPPR) for PD review and approval.
- It is the responsibility of the program director to review and approve the Institutional Certification. The PD should send any questions to the division or center's GPA[13] for clarification (see list in Supplement 2).
- The Program Director should work with the submitting PI to finalize the IC memo(s), ideally prior to funding. If the IC memo cannot be finalized prior to funding, a restricted award will be issued and the IC memo(s) must be approved by the end of the first year of funding.
- The approved IC memo(s) should either be attached to the greensheet or sent by the PD to the NCI Office of Grants Administration (OGA) for inclusion in the official grant file.
- The approved IC memo(s) should be sent to the designated GPA for upload to the dbGaP submission system.

***Intramural programs***

- IC memo(s) should be filled out and signed by the study PI and SD prior to data generation.
- For any study that requires consent of a human subject, the IRB should review the informed consent and protocol to assure that data sharing is appropriate and to interpret any data use limitations that may exist based on the language found in the consent. The Scientific Director, or delegate, should implement a process to make sure that adequate IRB review has occurred before signing the IC memo.
- In studies for which the Office of Human Subjects Research Protections (OHSRP) has provided an NIH Intramural investigator an exemption[14] the NIH Intramural investigator should ask the outside institution that is contributing the de-identified specimens/data to provide an institutional certification (IC) using the single site extramural IC memo document.
- The approved IC memo(s) should be sent to the designated GPA for upload to the dbGaP submission system.

## VII. Timelines

- Extramural program staff and intramural program leadership should monitor the progress of projects falling under the GDS policy throughout their life cycle to ensure that progress related to GDS is consistent with the DSP and be proactive in taking action to address any issues that arise.

[12] An Institutional Signing Official is generally a senior official at an institution who is credentialed through 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 NIH.

[13] GPA (http://gds.nih.gov/04po2_2GPA.html)

[14] The NIH investigator is getting coded samples and there is an identification agreement between the NIH investigator and the outside investigator providing the samples which basically assures that the NIH investigator will never receive identifying information

***Human Data Submission and Release***

Study Registration:

- Each division or center's GPA will register all studies with human genomic data that fall within the scope of the GDS Policy in dbGaP regardless of which NIH-designated data repository will receive the data.
- If an exception to the GDS policy has been granted by the NCI, the study will be registered in dbGaP in accord with this timeline, but data deposition will not be expected.
- The PD should ensure that the GPA has the Basic Study Information (including study description) (see example in Supplement 5) from the grantee in order to register the study in dbGaP when the data cleaning process begins.

Data Submission*:*

- So that a) submitted datasets are of the highest quality; b) submitted data are most useful for the secondary users; c) the likelihood that datasets will need revision post-submission is minimized; and d) efficiency of the submission process for PI, GDS staff, and the NCBI data curators is maximized; submission of data is generally expected once the data has been cleaned (e.g. the analytical dataset is finalized).
- Differences in this approach may occur depending on study type. For example, data generated by community resource projects (e.g., TCGA) may be required to be submitted on an accelerated timetable. Submitting PIs may also deposit datasets on an accelerated timetable if they so choose.

Data Release:

- Following data submission, the data may be held in an exchange area accessible only to the submitting investigators and collaborators for a period not to exceed six months. Following this period of exclusivity, or at the time of publication (whichever comes first), the data will be available for secondary research access without restrictions on publication.
- The PD or intramural GPA will determine if a shorter timeframe is warranted based on the publication status of the initial publication. At NCI's discretion, NCI may decide that data from community resource projects could be released earlier than 6 months after submission regardless of publication status.
- The PD or intramural GPA might also determine if a longer timeframe is warranted.

***Non-Human Data Submission and Release***

- Consistent with the 2004 Policy On Sharing Of Model Organisms For Biomedical Research[15], and the NIH GDS Policy, non-human data (including microbial data) from large-scale genomic projects for model organisms and, when appropriate, relevant phenotype data should be shared through openly accessible community resource data repositories[16] no later than the time of publication.
    - However, In many cases, NCI may decide that larger projects (*e.g.*, specific R01s involving over $500,000 in total costs in a single year for which robust data sharing is already expected[17] or consortia-based projects) or projects of high scientific priority (as determined by NCI leadership) will require pre-publication data submission and release.
    - Smaller-scale projects should share data through broadly accessible repositories no later than the time of publication, unless the investigator includes a justification in the data sharing plan submitted at the time of the funding request or through the appropriate intramural process demonstrating that data

[15] http://grants.nih.gov/grants/guide/notice-files/NOT-OD-04-042.html

[16] For example, the Gene Expression Omnibus (GEO), Sequence Read Archive (SRA), Trace Archive, Array Express, Mouse Genome Informatics (MGI), WormBase, the Zebrafish Model Organism Database (ZFIN), and GenBank.

[17] http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

sharing costs (*e.g.*, financial, time, personnel) outweigh the potential broad scientific value of the data and the NCI agrees that the data sharing costs outweigh the benefit.
    - In the case of de novo sequencing for non-human organisms, investigators who are submitting data prior to publication may request and NCI may agree to a holding period, not to exceed six months, during which the datasets will not be released for use by other investigators.

## VIII. Requests for Exception to the Policy

- The Institute recognizes that open or controlled access data sharing may not always be appropriate. In such rare cases, NCI will consider requests for an exception to usual data submission expectations.
- Submission of genomic data to an NIH data repository (e.g. dbGaP) may be precluded by various factors, such as international laws, limitations in the original informed consents, concerns about harms to individuals or groups, or other cases where expectations for data submission cannot be met.

***In cases where data submission to an NIH-designated data repository is not appropriate:***

1. Investigators should provide a justification for any data submission exceptions requested in the funding application or proposal. This justification must also include an alternative plan to share data through other mechanisms whenever possible.
2. NCI Divisions will consider the Statement of Scientific Merit and exception request at the time funding decisions are made and develop a Statement of Programmatic Priority.
3. Exception requests (including Statements of Programmatic Priority) will be reviewed by the Trans-NCI Data Sharing Working Group and a recommendation will be made.
4. NCI Scientific Program Leaders group (SPL) will review the exception request package. If it concurs with the recommendation from the Trans-NCI Data Sharing Working Group, the request will be sent to the NCI Director for signature.
5. Additionally, for intramural investigators requesting an exception, the request must be signed by the NIH Deputy Director for Intramural Research after the NCI Director.

***NCI will use the following criteria to assess the circumstances involved in the exception request:***

- Impact of data sharing compliance on scientific merit
- Uniqueness of the resource
- Value of the resource
- Regulatory considerations (e.g. limitations in the original informed consents)
- Ethical considerations
- NIH data sharing exception precedents
- Existence of an acceptable alternative data-sharing plan (ADSP)
    - Impact of ADSP on data re-use
    - Impact of ADSP on data discoverability
    - Burden
    - Feasibility
- In all cases where alternative data sharing plans are determined to be appropriate, information on how to request access to data and a basic summary of the study and study data will be listed in dbGaP (or other appropriate data repository).

## IX. Governance

- For extramural divisions the primary responsibility of implementing the GDS policy belongs with the Program Director.

- The ***Trans-NCI Genomic Data Sharing Working Group*** (WG) will provide on-going stewardship and leadership for Institute data sharing policies and their implementation. Specifically, the working group is charged with considering exception requests to any aspect of NIH or NCI data sharing expectations, addressing any on-going policy or implementation development needs, and adjudicating or interpreting any aspect of the NIH/NCI policies and practices.
    - The committee will report to the NCI Scientific Program Leaders on data sharing activities, resource needs, and compliance issues, and will consult with leadership as needed to carry out its responsibilities. This group will be constituted with representatives of each NCI Division or Center.
    - A primary deliverable from the WG is this framework document to guide and promote consistency in genomic data sharing policy implementation at the Institute. This framework document will include guidance for NCI staff as well as the extramural scientific community regarding expectations for implementation of the GDS policy (scope, timeline, data standards, sharing or consent exceptions to the policy). Note that this document will evolve with time.
    - The WG also sees the development of two central information hubs (cancer.gov and myNCI web pages) seen as essential to educating NCI staff and the extramural scientific community about the GDS policy and communicating NCI's expectations regarding the policy. The WG is working with the NCI Office of Communications and Public Liaison (OCPL) to create these sites as well as develop the content stored there. Links will be provided as soon as possible.
    - The WG also recognizes the need to develop education materials for both NCI staff and the extramural scientific community and perform on-going training.
    - Important future activities of the WG will include developing systems to track compliance as well as minimize the compliance burden.

***NCI Data Access Committees (DACs)***

- Charged with providing oversight and monitoring of data access activities and participant protection needs related to all NCI supported datasets (both intramural and extramural).
    - The NCI currently has several Data Access Committees (iNCI, eNCI and TCGA), however for the purposes of efficiency and decreased burden on DAC members, the NCI plans to merge these committees in the near future.
    - The DAC will be constituted with representatives of each NCI research Division or Center. Each Division director will nominate members for three-year terms with staggered rotations.

***Genomic Program Administrators (GPAs)***

- Function as a central point of coordination and information about NCI data sharing activities and implementation of NIH and NCI policies.
    - The NCI will have a GPA representing each division or center (Supplement 2).
    - The GPAs will be the point of contact for all staff within a division or center regarding the implementation of data sharing expectations in the extramural and intramural research programs.
    - The GPAs will work in concert with one another as well as the Trans-NCI Genomic Data Sharing Working Group on questions regarding implementation of Institute and NIH policies.
    - The GPAs will serve as a liaison to the NIH Genomic Data Sharing governance structure through the Technical Standards and Data Submission (TSDS) Steering Committee. Primary responsibility for representing NCI on the TSDS steering committee will fall to the GPAs from the extramural Division of Cancer Control and Population Sciences (DCCPS) and the intramural Division of Cancer Epidemiology and Genetics (DCEG).

**Supplement 1:**

**Examples of projects for which the NCI anticipates data sharing (*regardless of study design*) include, but are not limited to:**

| | # of Specimens | |
|---|---|---|
| | **Human *(including human cell lines)*** | **Model Organisms, Non-Human Cell Lines, Infectious Organisms** |
| **SNP array data from >500K single nucleotide polymorphisms (SNPs)** *(e.g., GWAS data)* | 1,000 | 500 |
| **DNA sequence data from < 100 genes or regions of interest** *(e.g., targeted sequencing)* | 1,000 | 500 |
| **DNA sequence data from ≥ 100 genes or regions of interest** *(e.g., targeted sequencing, whole exome sequencing, whole genome sequencing)* | 100 | 50 |
| **Genome-wide RNA sequencing (RNA-seq) data** *(e.g., transcriptomic data)* | 100 | 50 |
| **Genome-wide DNA methylation data** *(e.g., bisulfite sequencing data)* | 100 | 50 |
| **Genome-wide chromatin immunoprecipitation sequencing (ChIP-seq) data** *(e.g. transcription factor ChIP-seq, histone modification ChIP-seq)* | 100 | 50 |
| **Metagenome (or microbiome) sequencing data** *(e.g., 16S rRNA sequencing, shotgun metagenomics, whole-genome microbial sequencing)* | 100 | 50 |
| **Metatranscriptome sequencing data** *(e.g., microbial/microbiome transcriptomics)* | 100 | 50 |

**NOTE: The number of samples includes distinct individuals, species, strains, samples, treatments, time points, and tissues. For example, data from 25 patients at 4 time points after treatment would reach a 100-sample threshold, as would data from 50 tumor-normal comparisons.**

**Additionally**, individual NIH Institutes or Centers (IC) may choose on a case-by-case basis to apply the Policy to projects generating data on a smaller scale depending on the state of the science, the needs of the research community, and the programmatic priorities of the IC, therefore investigators should consult with appropriate NIH Program Officers or your intramural Scientific Director as early as possible.

**Examples of smaller-scale projects that the NCI would likely mandate data sharing for include, but are not limited to:**

- Projects examining rare cancers, rare-cancer-related outcomes, or rare cancer subtypes.
- Projects focusing on under-studied populations.

**Examples of Research outside the Scope of the GDS Policy:**

Examples of NIH-funded research or research-related activities that are outside the Policy's scope include, but are not limited to, projects that do not meet the criteria in the above examples and involve:

- Instrument calibration exercises.
- Statistical or technical methods development.

**Supplement 2:**

**Genomic Program Administrator (GPA):**

- Each NCI Division and Center will have a GPA who serves as the focal point of contact within that division (or center) with regard to GDS-related questions. A GPA is anticipated to:
  - Serve as a GDS policy resource for Program Directors within each division (who will be primarily responsible for the policy's implementation) and for intramural scientists generating relevant data.
  - Serve as a liaison between the Program Directors or IRP scientists and the National Center for Biotechnology Information (NCBI).
  - Work to coordinate activities across divisions and to promote transparency and consistency in GDS policy implementation at the Institute across both the intramural and extramural programs.
- In addition to the GPA, some NIH ICs also have a GPA assistant who works on implementing the policy. GPA and GPA assistant roles may vary across different ICs.

**NCI Genomic Program Administrators:**

| Division or Center | GPA | GPA assistant or back up |
|---|---|---|
| CCG | Jaime Guidry Auvil | Daniela Gerhard (GPA back-up) |
| CCR | Kathleen Calzone | Anjan Purkayastha (GPA assistant) |
| DCB | Jennifer Stasburger | Sean Hanlon (GPA back-up) |
| DCEG | Margaret Tucker | Geoff Tobias |
| DCCPS | Charlisse Caga-anan | Elizabeth Gillanders (GPA back-up)<br>Tiffany Green (GPA assistant)<br>Sharna Tingle (GPA assistant back-up) |
| DCP | Nada Vydelingum<br>Claire Zhu | |
| DCTD | Tamara Walton | |

**Supplement 3:**

**Resources for Data Standards**

The NIH National Center for Biotechnology Information (NCBI) provides general guidance for submitting data to NIH data repositories[18,19]. More specific instructions for data submission, including data standards, are available for a number of NIH repositories: Gene Expression Omnibus (GEO)[20] database of Genotypes and Phenotypes (dbGaP)[21] database of Short Genetic Variants (dbSNP)[22] GenBank[23] and Sequence Read Archive (SRA)[24]. Additional information or resources regarding standards for data and metadata will be included on the GDS website[25] as they become available and widely adopted by the research community.

**Guidance for Data Submission and Data Release**

Different data types undergo different levels of data processing, and the expectations for data submission and data release are based on those levels. Table 1 describes the expectations for each level. NIH will review these expectations at regular intervals, and will publish updates on the GDS website and the research community will be notified through appropriate communication methods (e.g., the *NIH Guide for Grants and Contracts*). Note that information necessary to interpret controlled-access genomic data, such as study protocols, data instruments, and survey tools, should be submitted to share on an unrestricted basis (i.e., through unrestricted access) concurrent with the relevant Level 1, 2, 3, or 4 genomic data.

[18] Submit data to NCBI. See https://submit.ncbi.nlm.nih.gov/
[19] How to Submit Data to NCBI. See http://www.ncbi.nlm.nih.gov/guide/howto/submit-data/
[20] GEO. Submitting Data. See http://www.ncbi.nlm.nih.gov/geo/info/submission.html
[21] Steps for dbGaP Study Registration, Submission, and Release of Data. See http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document_name=HowToSubmit.pdf
[22] Submission of Small Variations to dbSNP. See http://www.ncbi.nlm.nih.gov/projects/SNP/how_to_submit.html
[23] GenBank. How to Submit Whole Genome Shotgun (WGS) Genomes. See http://www.ncbi.nlm.nih.gov/genbank/wgs.submit
[24] Steps for SRA Submission. See http://www.ncbi.nlm.nih.gov/books/NBK47529/
[25] See http://gds.nih.gov/05rr.html

| Date Type By Data Processing Level[26] | Level 0 | Level 1[27] | Level 2 | Level 3 | Level 4 | Metadata |
|---|---|---|---|---|---|---|
| | Raw data generated directly from the instrument platform | Initial sequence reads, the most fundamental form of the data after the basic translation of raw input | Data after an initial round of analysis or computation to clean the data and assess basic quality measures | Analysis to identify genetic variants, gene expression patterns, or other features of the dataset | Final analysis that relates the genomic data to phenotype or other biological states | Information around the experiment or study |
| **SNP array data from >500K single nucleotide polymorphisms (SNPS)** *(e.g., GWAS data)* | submission not expected | .CEL\|.TXT\|.IDAT[28] | not applicable[29] | .TXT | .TXT[30] | Metadata around the experiment or study and annotations that are necessary to reproduce any published table or analysis must be included with genomic data submissions. In particular, data pertinent to the interpretation of genomic data—such as associated phenotype data (e.g., clinical information), exposure data, relevant metadata, and descriptive information (e.g., |
| **DNA sequence data from < 100 genes or regions of interest** *(e.g., targeted sequencing)* | submission not expected | | .BAM[31] | **Arrays:** .TXT | .TXT | |
| | | | | **NGS:** .MAF\|.VCF\|.PED[32] | .TXT | |
| **DNA sequence data from ≥ 100 genes or regions of interest** *(e.g., targeted sequencing, whole exome sequencing, whole genome sequencing)* | submission not expected | | .BAM | **Arrays:** .TXT | .TXT | |
| | | | | **NGS:** .MAF\|.VCF\|.PED | .TXT | |
| **Genome-wide RNA sequencing (RNA-seq) data** *(e.g., transcriptomic data)* | submission not expected | .FASTQ\|.SFF\|.HDF5\| Complete Genomics Native | submission not expected | **Arrays:** .TXT | .TXT | |
| | | | | **NGS:** .WIG\|.TXT[33] | .TXT | |
| **Genome-wide DNA methylation data** *(e.g., bisulfite sequencing data)* | submission not expected | | .BAM | **Arrays:** .TXT | .TXT | |
| | | | | **NGS:** .MAF\|.VCF\|.TXT\|.BED[34] | .TXT | |
| **Genome-wide chromatin immunoprecipitation sequencing (ChIP-seq) data** *(e.g. transcription factor ChIP-seq, histone modification ChIP-seq)* | submission not expected | | .BAM | **Arrays:** .TXT | .TXT | |
| | | | | **NGS:** .WIG\|.TXT\|.BED | .TXT | |
| **Metagenome (or microbiome) sequencing data** *(e.g., 16S rRNA sequencing, shotgun metagenomics, whole-genome microbial sequencing)* | submission not expected | | .BAM | **NGS:** .WIG\|.TXT | .TXT | |
| **Metatranscriptome sequencing data** *(e.g., microbial/microbiome transcriptomics)* | submission not expected | | .BAM | **NGS:** .WIG\|.TXT | .TXT | |

[26] The file formats, accepted by GEO and SRA, listed for each data type, at each data level, apply to both tissue and germline samples derived from humans (including human cell lines), model organisms, non-human cell lines and infectious organisms.

[27] Level 1 data submission is expected only for RNA-seq data generated from human samples. Leve1 data are not expected for any other data types for human samples. Level 1 NGS data may be submitted only for the de novo sequencing of non-human organisms for which Level 2 data will not be submitted. For array data (Expression, ChIP-chip, Array-CGH, SNP) may be submitted to GEO in various platform-dependent formats.

[28] Given the risk of personally identifiable information (PII) being embedded in .IDAT files, the submission of .IDAT files for human sample data will be decided on a case-by-case basis.

[29] Arrays do not produce alignment/assembly data

[30] The final analysis that relates genomic data to phenotype or other biological states may be stored as a text file.

[31] Expected Base and/or Mapping Quality Scores need to be established.

[32] The .MAF, .VCF and .PED file formats are used to list mutation data.

[33] The .WIG format can be used to annotate the sample coverage profile

[34] The .BED format can be used to annotate methylation and CHiP-peak profiles

**Supplement 4:**

## NCI GENOMIC DATA SHARING PLAN

Data produced through this award will be shared in a manner consistent with data-sharing under the *NIH Genomic Data Sharing Policy* (NOT-OD-14-124).

**Intramural Project (Z01), Grant, or Contract number (if available):** Enter project, grant or contract #

**Project Title:** Enter project title

**Principal Investigator:** Enter Principal Investigator

**Investigators Affiliation (Institution/Division/Program/Branch):** Enter Institution

**1.Data to be Shared (mark all that apply):**

| | | | |
|---|---|---|---|
| *Species:* | ☐ Human | ☐Rat | |
| | ☐Mouse | ☐C. elegans | |
| | ☐Drosophila | ☐Yeast: Indicate Species | |
| | ☐Bacteria: Indicate Species | ☐Other: Indicate Species | |
| *Sample Type:* | ☐Tumor Tissue | ☐Normal Tissue | ☐Blood |
| | ☐Buccal | ☐Urine | ☐Other: Indicate sample type |
| *Analyte Type* | ☐DNA | ☐RNA | ☐Other: Indicate analyte type |

*Genomic data (See NCI GDS framework for data sharing thresholds for each data type):*

| | |
|---|---|
| ☐ SNP array data from >500K single nucleotide polymorphisms (SNPS)<br>*e.g., GWAS data* | Enter # of Samples |
| ☐ DNA sequence data from < 100 genes or regions of interest<br>*e.g., targeted sequencing* | Enter # of Samples |
| ☐ DNA sequence data from ≥ 100 genes or regions of interest<br>*e.g., targeted sequencing, whole exome sequencing, whole genome sequencing* | Enter # of Samples |
| ☐ Genome-wide RNA sequencing (RNA-seq) data<br>*e.g., transcriptomic data* | Enter # of Samples |
| ☐ Genome-wide DNA methylation data<br>*e.g., Illumina 450k or other platforms, bisulfite sequencing data* | Enter # of Samples |
| ☐ Genome-wide chromatin immunoprecipitation sequencing (ChIP-seq) data<br>*e.g. transcription factor ChIP-seq, histone modification ChIP-seq* | Enter # of Samples |
| ☐ Metagenome (or microbiome) sequencing data<br>*e.g., 16S rRNA sequencing, shotgun metagenomics, whole-genome microbial sequencing* | Enter # of Samples |
| ☐ Metatranscriptome sequencing data<br>*e.g., microbial/microbiome transcriptomics* | Enter # of Samples |
| ☐ Other: List data types | Enter # of Samples |

*Phenotype data:*

Data pertinent to the interpretation of genomic data, including the minimal phenotype information needed to reproduce the primary analysis —such as associated phenotype data (e.g., clinical information), exposure data, relevant metadata, and descriptive information (e.g., protocols or methodologies used)—will be shared. Individual-level Phenotype data will include, at minimum:

| |
|---|
| Describe the covariate information you will include with your data submission |

**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[1] or controlled-access[2]. A list of relevant databases can be found at: http://gds.nih.gov/02dr2.html.

Repository: Enter repositories to which data will be deposited

Repository Accession Number (if known): Enter accession #

If human data, how will be data be made available?

○ Unrestricted-Access ○ Controlled-Access

**3. Data Submission Timeline:**
We will submit the genotype/sequencing and phenotype data after the genotyping/sequencing data have been cleaned (i.e. once the QA/QC is complete and the analytical dataset is finalized).

We understand that following data submission, the data may be held for a period not to exceed six months. Following this period of exclusivity, or at the time of publication (whichever comes first), the data will be available for secondary research access without restrictions on publication (i.e. there will be no publication embargo).

Date submission is expected (approximate): _____

**4. IRB Assurance of the Genomic Data Sharing Plan:**
Has an IRB or analogous review body reviewed the genomic data sharing aspects of your project? If not, provide a timeline for such review.

○ Yes

○ Not yet (enter date of expected review) _____

○ Not applicable (e.g. no human data)

**5. Appropriate Uses of the Data:**
The NIH promotes the broad and responsible sharing of genomic research for 'general research use'. However, NIH also recognizes that in some circumstances broad sharing may not be consistent with the informed consent of the research participants whose data are included in the dataset. A data use limitation (DUL) statement is a brief written description of limitations, if any, on the distribution and use of human data submitted to controlled-access NIH designated data repositories, such as the NIH database of Genotypes and Phenotypes (dbGaP). Limitations on the data use should be described in the Institutional Certification. NIH provides *Points to Consider in Developing Effective Data Use Limitations*.

How will data be shared?

○ Data will be made available for general research use

○ Data will be made available with the following limitation(s): Enter use limitations here

[1] Data publically available to anyone
[2] Data made available for secondary research only after investigators have obtained approval from NIH to use the requested data for a particular project

○ Data sharing is not appropriate, an exception is being requested (if selected complete 5a and 5b)

Exceptions to Submission:
Submission of genomic data to an NIH data repository (e.g. dbGaP) may be precluded by various factors, such as international laws, limitations in the original informed consents, concerns about harms to individuals or groups, or other cases where expectations for data submission cannot be met. The Institute recognizes that open or controlled access data sharing may not always be appropriate. In such **rare cases**, NCI will consider requests for an exception to usual data submission expectations.

5a. If submission of human data generated in the study would be not be appropriate because the Institutional Certification criteria cannot be met, the investigator should explain why (explanation subject to NIH review):

| Provide Explanation |
|---|

5b. Describe an alternative mechanism for data sharing. If the NCI 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:

| Enter alternative data sharing plan |
|---|

**6. Approvals**

Principal Investigator: _____ Date: _____

Scientific Director, or designee
(Intramural only): _____ Date: _____

**Supplement 4:**

# dbGaP Basic Study Information

| **Basic Study Information** | | |
|---|---|---|
| Study Name: | | |
| Institute(s) or Center(s) supporting the study: | | |
| Estimated number of study participants: | | |
| Principal Investigator: | Name: | Email: |
| PI Assistant/Data Submitter: | Name: | Email: |

| **Data Types To Be Submitted (check all that apply)** | |
|---|---|
| **General:**<br>☐Individual Phenotype<br>☐Individual Genotype<br>☐Individual Sequencing<br>☐Supporting Documents<br>☐Metagenomic<br>☐Protomic/Metabolomic<br>☐Images | **Sample Types:**<br>☐Germline<br>☐Tumor/Normal<br>☐DNA<br>☐RNA<br>☐Mitochondria<br>☐Microbiome<br>☐From Repository |
| **Array Data:**<br>☐SNP Array<br>☐Expression Array<br>☐Methylation Array | **Genotypes:**<br>☐Array derived Genotypes<br>☐CNV calls from microarray<br>☐CNV calls derived from Sequencing<br>☐Genotype calls derived from Sequence<br>☐Somatic SNV (.MAF)<br>☐Array CGH CNVs |
| **Sequencing:**<br>☐Whole Genome<br>☐Whole Exome<br>☐Targeted Genome<br>☐Targeted Exome<br>☐Whole Transcriptome<br>☐Targeted Transcriptome<br>☐Epigenomic Marks<br>☐Sanger<br>☐16S rRNA | **Analyses:**<br>☐Association/Linkage Results<br>☐Array derived Expression<br>☐RNASeq derived Expression<br>☐Array derived Methylation |

| **Policy Information** | |
|---|---|
| Acknowledgement statement to be used by approved users:<br><br>*Example: The XYZ study was supported by the Intramural Research Program of the National Cancer Institute, National Institutes of Health, Department of Health and Human Services including Contract No. HHS1234. The datasets have been accessed through the NIH database for Genotypes and Phenotypes (dbGaP). A full list of acknowledgements can be found in the supplementary note (John Doe et al., PMID: 12345678). Please cite this publication in all oral or written presentations, disclosures, or publications in which these data were used.* | |
| IC-specific Terms of Data Access (if applicable): | |