Supplemental Information to the National Institutes of Health Genomic Data Sharing Policy

Overview

This document provides additional guidance on the scope of the Genomic Data Sharing (GDS) Policy[1] (i.e., examples of research that do and do not fall within the Policy's scope), expected use of data standards, and expectations for data submission and release.

Guidance Regarding Scope of the GDS Policy

Examples of Research within the Scope of the GDS Policy

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 genome-wide association studies (GWAS), single nucleotide polymorphisms (SNP) arrays, and 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).[2] Examples of such research include, but are not limited to, projects involving:

- Sequence data from more than one gene or region of comparable size in the genomes of more than 1,000 human research participants.
- Sequence data from more than 100 genes or region of comparable size in the genomes of more than 100 human research participants.
- Data from 300,000 or more variant sites in more than 1,000 human research participants.
- Sequence data from more than 100 isolates from infectious organisms.
- Sequence data from more than 100 metagenomes of human or model organism microbiomes.
- Sequence data from more than 100 metatranscriptomes of human or model organism microbiomes.
- Whole genome or exome sequence data of more than one model organism species or strain.
- Comprehensive catalog of transcripts and non-coding RNA from one or more model organism species or strains.
- Catalog of more than 100,000 single nucleotide polymorphisms (SNPs) from one or more model organism species or strains.
- Comparisons of genome-wide methylated sites across more than 10 cell types.
- Comparisons of differentially methylated sites genome-wide at single-base resolution within a given sample (e.g., within the same subject over time or across cell types within the same subject).

[1] See https://osp.od.nih.gov/wp-content/uploads/NIH_GDS_Policy.pdf..

[2] Smaller project sizes may have data that a particular NIH funding IC would find valuable and expect to be submitted, therefore investigators should consult with appropriate NIH Program Officers as early as possible.

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.
- Use of genomic data for control purposes, such as for assay development.

Resources for Data Standards

The NIH National Center for Biotechnology Information (NCBI) provides general guidance for submitting data to NIH data repositories.[3,4] More specific instructions for data submission, including data standards, are available for a number of NIH repositories: Gene Expression Omnibus (GEO),[5] database of Genotypes and Phenotypes (dbGaP),[6] database of Short Genetic Variants (dbSNP),[7] GenBank,[8] and Sequence Read Archive (SRA).[9] Additional information or resources regarding standards for data and metadata will be included on the NIH Office of Science Policy website[10] 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 and text below it describe the expectations for each level. NIH will review these expectations at regular intervals, and will publish updates on the NIH Office of Science Policy 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 2, 3, or 4 genomic data.

[3] Submit data to NCBI. See https://submit.ncbi.nlm.nih.gov/.
[4] How to Submit Data to NCBI. See http://www.ncbi.nlm.nih.gov/guide/howto/submit-data/.
[5] GEO. Submitting Data. See http://www.ncbi.nlm.nih.gov/geo/info/submission.html.
[6] 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.
[7] Submission of Small Variations to dbSNP. See http://www.ncbi.nlm.nih.gov/projects/SNP/how_to_submit.html.
[8] GenBank. How to Submit Whole Genome Shotgun (WGS) Genomes. See http://www.ncbi.nlm.nih.gov/genbank/wgs.submit.
[9] Steps for SRA Submission. See http://www.ncbi.nlm.nih.gov/books/NBK47529/.
[10] See https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing/.

Table 1: Expectations for Data Submissions and Release Based on Processing Level

Level	General Description of Data Processing	Example Data Types	Data Submission Expectation	Data Release Timeline
0	Raw data generated directly from the instrument platform	Instrument image data	**Human data:** Not expected. **Non-human data:** Not expected.	**Human data:** NA. **Non-human data:** NA.
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, Array CGH	**Human data:** Not expected. **Non-human data:** Not expected, except for de novo sequence data (unless it is included with Level 2 aligned sequence files). Submission of de novo sequence data is expected no later than the time of initial publication.	**Human data:** NA. **Non-human data:** No later than the time of initial publication; an earlier release date may be designated for certain data types or NIH projects.
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	**Human data:** Project specific; after data cleaning and quality control, which is generally within 3 months after data have been generated. **Non-human data:** Data submission is expected no later than the time of initial publication; an earlier submission date may be designated for certain data types or NIH projects.	**Human data:** Up to 6 months after data submission is initiated or at the time of acceptance of initial publication, whichever occurs first. **Non-human data**: No later than the time of initial publication; an earlier release date may be designated for certain data types or NIH projects.
3	Analysis to identify genetic variants, gene expression patterns, or other features of the dataset	SNP or structural variant calls, expression peaks, epigenomic features	**Human data:** Project specific; after cleaning and quality control, which is generally within 3 months after data have been generated. **Non-human data:** Data submission is expected no later than the time of initial publication; an earlier release date may be designated for certain data types or NIH projects.	**Human data:** Up to 6 months after data submission is initiated or at the time of acceptance of initial publication, whichever occurs first. **Non-human data:** No later than the time of initial publication; an earlier release date may be designated for certain data types or NIH projects.
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	**Human data:** Data submitted as analyses are completed. **Non-human data:** Data submission is expected no later than the time of initial publication.	**Human data:** Data released with publication. **Non-human data:** No later than the time of initial publication.

Level 0 Data: These data are the raw images and generally have limited value to secondary data users. NIH policy does not expect submission of these data.

Level 1 Data: These data are the initial sequence reads and generally have limited value to secondary data users. NIH policy does not expect submission of these data, except for de novo sequence data from non-human organisms (unless it is included with Level 2 aligned sequence files). 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.

If investigators choose to submit level 1 human data to an NIH-designated data repository, it is the submitting institution's responsibility to protect participant privacy by ensuring that data submission is consistent, as appropriate, with all applicable national, tribal, and state laws and regulations as well as relevant institutional policies, and the GDS Policy.

Level 2 Data: These data constitute a computational analysis in the form of higher order assembly or placement of the sequencing reads on a reference template. The level 2 file comprises the reads "piled" on a reference genome. A submission would be a file (e.g., binary alignment matrix (BAM) files) that contains 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.

Preparation of level 2 data generally requires substantial data cleaning, analysis, and quality checks related to both breadth of coverage of the targeted region and accuracy of assembly. Sufficient time will be allowed to clean the data by removal of extraneous or poor-quality sequence, complete quality-control analyses, and generate the assembly, up to the coverage and quality thresholds specified by a project or investigative team. It is anticipated that this work could generally be completed within three months, and data submission would follow shortly thereafter, but this may vary depending on the data type or specific program design.

After submission of human data begins, 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, 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 three months after data cleaning begins.

Level 3 Data: These data include analyses 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 for human data 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 unrestricted- or controlled-access mechanisms as appropriate and without publication limitations.

Level 4 Data: These data constitute 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.