



How to upload raw human genetic data:

Overall introduction

Below, we will provide a step-by-step guide on uploading raw human genetic data using [dbGaP](#) (Database of Genotypes and Phenotypes). This database is a resource developed to archive and distribute data and results from studies exploring the relationship between genotype and phenotype in humans.

Note that this kind of data requires **control access**: Controlled access means that the raw data cannot be freely downloaded by anyone. Instead, access is restricted to approved researchers who apply and justify their intended use. This is done to protect participant privacy, ensure ethical use aligned with the original consent given by participants, and prevent misuse of sensitive data.

For Example:

If you want to access individual-level genotype data from [dbGaP](#), you must:

1. Submit a Data Access Request.
2. Get approval from a Data Access Committee (DAC).
3. Agree to follow certain rules.

The following guide is based on the National Library of Medicine guide:

<https://www.ncbi.nlm.nih.gov/gap/docs/submissionguide/#asc>

If you wish to upload a different kind of data, the process will likely be much simpler. To get a suggestion of the best database for your data, use [this link](#). Then, using your portal ([My submissions](#)), use the type of data you will upload:

Submission Portal

Your submissions

Start a new submission

- [GenBank](#)
- [Sequence Read Archive](#)
- [Genome](#)
- [TSA](#)

- [BioProject](#)
- [BioSample](#)
- [Supplementary Files](#)
- [API](#)

- [dbGaP](#)
- [GTR](#)
- [ClinVar](#)

Repository: [dbGaP](#)

Make sure you have an NCBI account and have set up your profile information. Only then will you be able to upload to this repository.

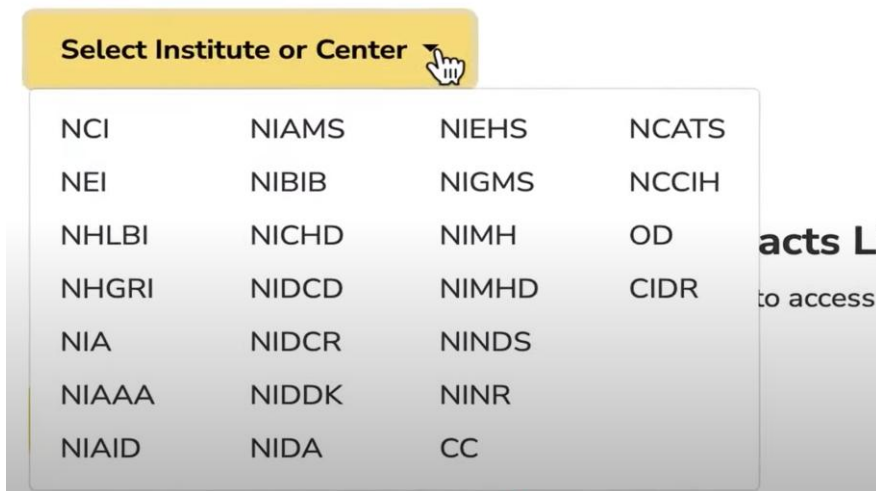
Register your study

Link to explanatory YouTube guide: <https://www.youtube.com/watch?v=P79c3gAWgP4>

1. Identify a Genomic Program Administrator (GPA):

Investigators cannot directly register a study on their own, therefore, they need to identify who is their GPA, based on the funding of the study. GPAs are designated by NIH institutes or centers:

Choose the IC that supports your research to find their GPA(s)



The image shows a screenshot of a web interface with a yellow button labeled "Select Institute or Center" with a hand cursor icon. Below the button is a dropdown menu displaying a list of NIH institutes and centers in a grid format. The list includes:

NCI	NIAMS	NIEHS	NCATS
NEI	NIBIB	NIGMS	NCCIH
NHLBI	NICHD	NIMH	OD
NHGRI	NIDCD	NIMHD	CIDR
NIA	NIDCR	NINDS	
NIAAA	NIDDK	NINR	
NIAID	NIDA	CC	

Once the correct GPA has been identified, you will be required to give basic study information, including:

- Number of participants
- Data types
- Consent groups- These are categories that describe what participants have agreed to regarding data use. For example, some participants may have consented only to disease-specific research, while others may have agreed to general biomedical research. You'll need to specify which participants fall into which group.
- Institutional Certifications: These are formal documents provided by your institution (often through the IRB) confirming that the data you're submitting was collected and will be shared in accordance with participants' informed consent and ethical standards. It also certifies that the data use aligns with institutional and federal policies.

Once all these are submitted, the GPA will process the request and initiate the registration of your study. Then, the principal investigator and the assistant will receive an email to approve the registration.

Submitting your data

Link to explanatory YouTube guide: <https://www.youtube.com/watch?v=L1jOi0w9fwg>

1. **Study Data Outline:** The first step is to submit the data outline in the [Submission Portal](#) in order to assert what data types will be submitted and released for the current study. Upon completion, a dbGaP study access point/reference number will be provided (e.g., phs#####.v#.p#). Make sure you only submit data that is mentioned in the outline.

To edit the Study Data Outline, go to the upper right section of the Submission Portal and click on **"Study Data Outline."**

2. **Study Config:** Fill out the online study configuration form, which collects detailed information about your study, such as data types, methods, key findings, inclusion/exclusion criteria, study history, references, and acknowledgments.

To complete the study configuration, visit your study's dbGaP Submission Portal at <https://submit.ncbi.nlm.nih.gov/dbgap/>.

- Click **"Create"** to start a new study configuration, or **"Edit"** to update an existing one.
- When finished, click **"Submit"** to return to your study's main Submission Portal page.
- To view a preview of your submission, select **"Preview Study Report Page."**

3. What can be published?

Below is a checklist of all possible data files. The links will guide you to the NCBI site with an example for each of the data files. The more you adhere to the guidelines, the easier it will be for your upload request to be processed and accepted. Below is a short list are brief explanations of what each of these data files refer to.

I. Phenotype Dataset (DS) and Data Dictionary (DD) files

- a. (1) [Subject Consent DS and DD](#)
- b. (1) [Subject Sample Mapping \(SSM\) DS and DD](#)
- c. (1) [Pedigree DS and DD](#)
- d. (1 or more) [Subject Phenotypes DS and DD](#)
- e. (1 or more) [Sample Attributes DS and DD](#)

- f. (1 or more) [Linking Subject/Sample IDs to samples in other NCBI databases DS and DD](#)

- II. [Molecular Data](#)
- III. [Sequence Data](#)
- IV. [Association Analysis Data](#)
- V. [Study Documents](#)
- VI. [Medical Images](#)

I. **Phenotype Datasets (DS) and Data Dictionaries (DD)**

- DS (Dataset): A table with your study's data—like participant IDs, clinical info, or consent group.
- DD (Data Dictionary): A file that explains each column in the dataset—what it means, what values are allowed, and the data type.

You need to submit both the dataset (DS) and its matching dictionary (DD) for each type of data.

What may be submitted?

- Subject Consent DS/DD
If your study includes participants who gave consent, you must include a dataset showing each participant's consent group.
- Subject Phenotypes DS/DD
If you have participant-level data (like age, sex, diagnosis, or exposures), submit one or more datasets with that information and the corresponding dictionaries.
- Sample Mapping & Attributes DS/DD
If you have molecular or sequencing data (like arrays, methylation, BAM/FASTQ files), include:
 - A mapping file linking participants to their samples
 - Sample attribute files describing those samples
- Pedigree DS/DD
If your study includes related individuals (like twins or family members), submit a pedigree file showing those relationships.

- **Linking DS/DD to Public Databases**

If your samples are also submitted to public databases (like GEO or GenBank), provide a dataset linking your subject/sample IDs to the database accessions.

- If you only have project-level accessions, mention this in the Study Config form and mark "no" for the question about individual ID linking.

II. Molecular Data

Includes data like: GWAS results, SNP arrays, Imputed genotypes, Transcriptomics, Epigenomics, Gene expression, Expression counts, Variant calls from WGS, WXS, or targeted sequencing

Note: This does **not** include raw sequencing files (BAM, CRAM, FASTQ)—those are handled separately.

III. Sequence Data

High-throughput sequence data (such as WGS, WXS, or RNA-Seq) in BAM, CRAM, or FASTQ format will include:

1. Sequence metadata file that defines the sequences to be submitted
2. The actual sequence

IV. Association Analyses

Any aggregated genomic-level data. Because this is summary-level data, there is no subject consent tied to these analyses.

V. Study Documents

Any consent forms, protocols, questionnaires, etc., that correspond to the data.

VI. Medical Images

Any CT scans, eye images, etc.

4. Final steps: review and preprocessing

Your uploaded data will undergo both automated and manual quality control checks. The system will automatically flag any common errors and notify you if corrections are needed. Since this process can be demanding, ensure that your data follows the required format and remains consistent across all submitted files.

In this [link](#) are several questions that are frequently asked during this process.