**Getting a study released in dbGaP**

Please read the following instructions before submitting to dbGaP. This file explains dbGaP’s editorial criteria, and how studies are handled by dbGaP from submission to the final release of the public study page and Authorized Access data. The dbGaP Authorized Access is the management portal for individual-level data. This site can be used to submit a data access request, manage access requests, and download approved data sets.

**Submission:**

Please submit all files through the dbGaP Submission Portal. dbGaP is unable to accept individual level data via email due to security reasons. We will send you an email to access your Submission Portal account. Please also include preprints of unpublished articles for your study if applicable; they will be used internally to better enable dbGaP to understand and represent your study.

**Loading Process/Quality Control Checks/Resubmissions:**

In order to correctly load your study into the dbGaP database, dbGaP will run several quality control (qc) checks on the data and data dictionaries and will report back any errors, such as HIPAA identifiers, inconsistencies in subject consent, gender, pedigree, affection status and formatting errors. We will provide you at various points with a preview site of your study as it might appear on the final public dbGaP page for you to review. Resubmissions are expected. When resubmitting to the dbGaP Submission Portal, please only submit files that are new or have been updated. Please do not resubmit files that have no changes.

**Release Announcement:**

The release occurs approximately 8 weeks after dbGaP receives data that is without error. If errors are uncovered through qc checks, please expect the processing time to increase. Once dbGaP and the submitter approve of the presentation of the study, it will be released to the [public dbGaP webpage](http://www.ncbi.nlm.nih.gov/gap), [public FTP site](ftp://ftp.ncbi.nlm.nih.gov/dbgap/studies/) and to the [Authorized Access](https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login) portal, where approved investigators can access the individual level phenotype data and genetic data.

**Quick Start:**

1. Fill out the [Study Configuration](#Study_Configuration) file
2. Select [Study Logo and URL](#Study_Logo)
3. Create Data Files and Data Dictionary Files
   1. Generate [Subject Phenotype Data Files](#Phenotype_DS)
   2. Generate [Sample Attributes Data Files](#Phenotype_DS)
   3. Generate [Phenotype/Attributes Data Dictionary](#Phenotype_DD)
   4. Create a [Subject Consent Data File](#Subject_Consent_DS)
   5. Create a [Subject Consent Data Dictionary File](#Subject_Consent_DD)
   6. Create a [Subject Sample Mapping Data File](#Subject_Sample_DS)
   7. Create a [Subject Sample Mapping Data Dictionary File](#Subject_Sample_DD)
   8. Create a [Pedigree Data File](#Pedigree_DS)
   9. Create a [Pedigree Data Dictionary File](#Pedigree_DD)
4. Collect the [Study Documents](#Study_Documents)
5. Complete [Molecular Data](#Molecular_Data) (ex. genotype, CNV, Trace, GEO, GENBANK, SRA) submission
6. Complete [Association Analysis Data](#Association_Analysis_Data)
7. Submit Study
8. **Study Configuration**

See an example of a study report page generated from the study configuration file: <http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?id=phs000001>

The Study Configuration File contains information that should be presented on the study- or sub-study report page:

## dbGaP study name

## Study report name

## Description

## Study URL

## Study Logo URL (See #3 Study Logo)

## Study Type

## Molecular Data

## Phenotype term(s)/Disease name(s) linked to Entrez MeSH

## Gene(s)

## Inclusion/Exclusion criteria for participants

## History of Study

## Relevant Publications

## Attribution

1. Grouping dbGaP studies “By Disease”

|  |  |
| --- | --- |
| Action | * Open **1\_dbGaP\_StudyConfig.docx** in the directory Phenotype\_Data, and follow the instructions to fill out the study configuration file. |
| Accepted Format | txt, doc |

1. **Study Logo**

A Study Logo is a high quality study image that is at least 200px by 200px. The study logo(s) appear on the bottom of the study report page (see <http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000001.v1.p1>). More than one logo can be submitted. Each logo must be submitted with an associated url, so dbGaP can link the url to the logo.

|  |  |
| --- | --- |
| Action | * Re-size logos if necessary. * Open **1\_dbGaP\_StudyConfig.docx** in the directory Phenotype\_Data, and fill in the two sections titled Study URL and Study Logo. |
| Accepted Format | jpg, png, tiff, gif, bmp |
| Common errors that require resubmission | * Logos smaller than 200px by 200px. |

1. **Data File and Data Dictionary File**

To introduce the dbGaP Data Files and Data Dictionary Files, below are dbGaP terms that you will need to know.

**What is a dbGaP Subject (SUBJECT\_ID)?**

A dbGaP Subject is defined as a single human person/individual/patient that arises from a single germline. Each subject should be submitted with a single, unique, de-identified subject ID. Subject IDs should be an integer or string value. Integers should not have zero padding. Only the following characters can be included in the ID: English letters, Arabic numerals, period (.), hyphen (-), underscore(\_), at symbol (@), and the pound sign (#). Once a variable name for the subject ID has been chosen, please use the same variable name throughout all the phenotype files for consistency. For example, please do not use SUBJECT\_ID in one file and INDIVIDUAL\_ID in another file. In addition to the submitted subject ID, dbGaP will assign a dbGaP subject ID that will be included in the final dump files along with the submitted subject ID.

***IMPORTANT NOTE ON SUBJECT DE-IDENTIFICATION*:**

**The subject (person) identifier submitted to dbGaP should NOT be linked directly to personal identifiers in any study records. Submitted subject identifiers should be two steps removed from any personal information.**

**For example:**

**Step one: Personal Information 🡪 Study person ID**

**Step two: Study person ID 🡪 Subject ID submitted to dbGaP.**

**Submitted Subject ID may be randomly assigned or a consecutive number that provides no information that would link to subject identity (i.e., the submitted Subject ID should not be based on the study person ID or on any personal identifiers such as subject’s birth date, health record number, or name).**

**\*De-identification applies to all IDs submitted in the data files.**

**HIPAA: Sensitive, personally identifying information must be removed prior to submission. Sensitive information also includes names, cities, all elements of dates (including months and days), telephone numbers, social security numbers, and any other unique identifying number, characteristic or code.**

**What is a dbGaP Sample (SAMPLE\_ID)?**

A dbGaP Sample is defined as the IDs of the final preps submitted to dbGaP by a genotyping center, to the SRA group by a sequencing group, or to a NCBI resource, such as GEO or BioSample. A single subject may be mapped to multiple samples, but a single sample should not be mapped to multiple subjects. Each sample should be submitted with a single, unique, de-identified sample ID. Sample IDs should be an integer or string value. Integers should not have zero padding. Only the following characters can be included in the ID: English letters, Arabic numerals, period (.), hyphen (-), underscore(\_), at symbol (@), and the pound sign (#). Once a variable name for the sample ID has been chosen, please use the same variable name throughout all the phenotype files for consistency. For example, please do not use SAMPLE\_ID in one file and SAMPLE\_NAME in another file. In addition to the submitted sample ID, dbGaP will assign a dbGaP sample ID that will be included in the final dump files along with the submitted sample ID. For example, if one subject (SUBJECT\_ID) provided one sample, and that sample was processed to generate 2 sequencing runs or 1 sequencing and 1 genotyping array run, the data file would show two rows, both using the same subject ID, but having 2 unique sample IDs.

**What is a dbGaP Variable?**

A dbGaP Variable is defined as the variable name and associated column of data in a data table. The variable’s metadata, such as the variable name, description, units, type, and encoded values are defined in its respective [Phenotype Data Dictionary Files](#Phenotype_DD).

**What is a dbGaP Data File?**

Data Files are rectangular tables of data values. There are 5 types of data files:

1. Subject Phenotype Data File(s) – 1 or more file per study
2. Sample Attributes Data File(s) – 1 or more file per study
3. Subject Consent File – 1 file only per study
4. Subject Sample Mapping File – 1 file only per study
5. Pedigree File – 1 file only per study if applicable

In the data files, each column in the data table is a single phenotypic variable. Row #1 (column headers) of a data file will only contain the variable names.

In the data files, each row contains phenotypes of one Subject or attributes of one Sample. After the first row (for column headers of variable names) each subsequent row of a data file should reflect one subject or sample depending on the type of file.

**What is a dbGaP Phenotype Data Dictionary File?**

Data Dictionary (DD) Files are tables that define and describe the variables in the data files. Each Data file must be submitted with its matching DD file.

**a. Subject Phenotype (1 or more file per study) and Sample Attribute Data Files (1 or more file per study)**

Subject Phenotype or Sample Attributes Data Files contain measured and/or descriptive traits per individual subject or sample respectively. Every subject (SUBJECT\_ID) that appears in a Subject Phenotype Data File must belong to a consented subject (to allow his/her phenotypes to be used by approved Authorized Access Users) and every sample (SAMPLE\_ID) that appears in a Sample Attribute Data File must belong to a consented subject. This means that every subject listed in a Subject Phenotype Data File must be listed in the Subject Consent Data File. Every sample listed in a Sample Attribute Data File must map to a subject that is consented in the Subject Consent Data File. Additionally, the Subject and Sample IDs used in the phenotype/attribute data files must be the primary IDs listed in the Subject Consent File and Subject Sample Mapping File. The IDs should not be the alias IDs that may be included under variables “SOURCE\_SUBJECT\_ID” of the Subject Consent Data File or “SOURCE\_SAMPLE\_ID” of the Subject Sample Mapping Data File.

* **Subject Phenotype Data Files**:

The Subject Phenotype Data File contains phenotype data pertaining to the consented study subjects (click here to see definition of [dbGaP Subject](#dbGaP_Subject)). To see an example of a Subject Phenotype Data File, open **2a\_dbGaP\_SubjectPhenotypesDS.txt** in the directory Phenotype\_Data using Excel. A few examples of subject phenotype are gender, ethnicity, medical history, physical observations and laboratory measurements. The primary IDs in this file are Subject IDs (SUBJECT\_ID).

* **Sample Attribute Data Files**:

The Sample Attribute Data File contains attributes of study samples (click here to see definition of [dbGaP Sample](#dbGaP_Sample)). To see an example of a Sample Attribute Data File, open **3a\_dbGaP\_SampleAttributesDS.txt** in the directory Phenotype\_Data using Excel.

For all Sample Attribute Data Files, the primary IDs are Sample IDs (SAMPLE\_ID). Include minimally the four sample attributes below. These four sample attributes along with the subject’s sex value will also be displayed on the NCBI BioSample webpage (<http://www.ncbi.nlm.nih.gov/biosample/>).

* + **Body Site** – the body part from which the sample is collected (ex. skin, breast, peripheral blood and inner oral cavity). If the sample is from a xenograft, feel free to rename the variable name.
  + **Analyte Type** – the analyte type of the sample (ex. DNA and RNA)
  + **Histological Type** – the type of sample, cell or tissue type or subtype of the sample (ex. melanoma, adenocarcinoma, melanocytes, keratinocytes, buccal cells and embryonic stem cells)
  + **Is Tumor** – the tumor status of the sample. The values can be binary such as yes/no or the values can be a string such as normal, tumor, benign tumor, malignant tumor, primary and metastatic tumor. For non-cancer studies, the values in IS\_TUMOR should be “no” if known and “unknown” if not known.

Please note that the example file we have included is based on a cancer study and the variables that are included would be useful for cancer studies. However, if the study is not a cancer study, additional sample attribute variables should also be included if they can provide greater understanding to the study. A few examples are sample extraction method, if the sample genome is amplified, sample plate or well number, sample QA results, and sample affection status (ex. psoriatic skin sample vs. non-psoriatic skin sample from a case subject who has psoriasis).

|  |  |
| --- | --- |
| Action | * Determine whether you are creating Subject Phenotype Data Files, Sample Attribute Data Files or both types. Create the data files so that Subject Phenotype Data Files have de-identified Subject IDs (SUBJECT\_ID), and Sample Attribute Data Files have de-identified Sample IDs (SAMPLE\_ID). * If the file is a Subject Phenotype File, check that every row has a Subject ID in the Subject ID column. If the file is a Sample Attributes File, check that every row has a Sample ID in the Sample ID column. There can only be one main ID column per data file and every value for the ID column must be filled. * Check for HIPAA values and remove any identifying information. * In case of repeated measures, where the same Subject ID or Sample ID is listed in multiple rows, a unique key [variable combination] needs to be submitted; see Phenotype Data Dictionary below, “UNIQUEKEY”). * Check formatting. For tab-delimited txt files, open in a txt editor, and remove any unintended quotes (“””). Remove empty rows between data values. |
| Accepted Format | * Each data worksheet must be in a separate file. Please do not submit multiple data worksheets per file and especially please do not submit the data and data dictionary as different worksheets in the same file. * Tab-delimited txt\*, xls and xlsx only. * \*Tab-delimited txt files are preferable. The final dump files provided to Authorized Users of the study will be in the tab-delimited txt format.   (File names should not contain special characters)   * All data file names must be submitted without spaces, hyphens, brackets, periods, or forward (/) or backward slashes (\). |
| Common errors that require resubmission | * Subject or Sample IDs are not de-identified. * IDs include characters other than the following approved characters: English letters, Arabic numerals, period (.), hyphen (-), underscore(\_), at symbol (@), and the pound sign (#). * Data files contain sensitive, personally identifying information. * Subject IDs are missing in a Subject Phenotype Data File or Sample IDs are missing in a Sample Attributes Data File. * Subject IDs and/or Sample IDs are not the primary IDs, but alias IDs listed in either “SOURCE\_SUBJECT\_ID” of the Subject Consent Data File or “SOURCE\_SAMPLE\_ID” of the Subject Sample Mapping Data File. * Converted tab-delimited txt files contain multiple unintended quotes (“””). * Data tables contain information that belongs to the data dictionary (e.g., variable descriptions, comments, notes). * CSV files have commas embedded in data values. |

**Subject Phenotype and Sample Attribute Data Dictionary Files – 1 data dictionary for every data file**

Each Data Dictionary (DD) should contain descriptions/definitions/algorithms of every variable found in its complementary Data File (i.e., Phenotype Data File and Sample Attributes Data File). List one variable and description per row. Variable names in the DD file must have the same spelling, spacing, syntax to those in the data file and preferably be unique throughout the study. The column headers “VARNAME and VARDESC” are required; other headers may be included whenever applicable.

To see an example of a Subject Phenotype Data Dictionary File that matches to **2a\_dbGaP\_SubjectPhenotypesDS.txt**, open **2b\_dbGaP\_SubjectPhenotypesDD.xlsx** in the directory Phenotype\_Data using Excel.

To see an example of a Sample Attribute Data Dictionary File that matches to **3a\_dbGaP\_SampleAttributesDS.txt**, open **3b\_dbGaP\_SampleAttributesDD.xlsx** in the directory Phenotype\_Data using Excel.

|  |  |
| --- | --- |
| **Column Headers** | **Description** |
| **VARNAME\*** | **Var**iable **name**. The VARNAME must not contain forward (/) or backward slashes (\) or commas (,). |
| **VARDESC\*** | **Var**iable **desc**ription. For variable descriptions referring to a gene, please identify the particular gene by adding the Entrez “GeneID.” See example of the KRAS gene: <http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=search&term=kras> |
| **DOCFILE** | Document name associated with the variable. To list multiple documents, add a semicolon (;) between documents. The documents listed should be submitted to dbGaP. |
| **TYPE** | Data value type: **integer** (1,2,3,4,…), **encoded value** (integers or strings are coded for non-numerical meaning, ex. 1=Control; 2=Case, see VALUES), **decimal** (0.5,2.5,…), **string** (African American, Asian, Caucasian, Hispanic, Non-Hispanic). For **mixed values** (any combination of string, integers, decimals and /or encoded values) in a single data column, list all types present. |
| **UNITS** | Units of measurement of variable (if applicable) |
| **MIN** | The logical minimum value of the variable. If a separate code such as -1 is used for a missing field, this should not be considered as the MIN value. |
| **MAX** | The logical maximum value for the variable. If a separate code such as 9999 is used for a missing field, this should not be considered as the MAX value. |
| **RESOLUTION** | (Optional) Measurement resolution – the number of decimal places to which a measured value is presented in the data. For example, in 54.321 the resolution is 3. |
| **COMMENT1**  **COMMENT2** | (Optional) Additional information not included in the VARDESC that will further define the variable. If additional comments are needed beyond COMMENT2, insert new columns (COMMENT3, COMMENT4, etc.) before the column “ORDER.” |
| **VARIABLE\_SOURCE** | (Optional) Source of controlled vocabularies. Ex. PhenX, MeSH, SNOWMED, NCI, etc. If there is no match, leave blank. |
| **SOURCE\_VARIABLE\_ID** | (Optional) A unique identifier from the VARIABLE\_SOURCE or a unique text concept/term from various controlled vocabularies. |
| **VARIABLE\_MAPPING** | (Optional) For example, a variable from the source could be Identical, Related, or Comparable. |
| **UNIQUEKEY** | (Longitudinal datasets only) Unique key is a combination of variables that is designed to uniquely identify a row in the longitudinal dataset. Mark “X” for variables that constitute the unique keys, and leave other values blank. Ex. SUBJECT\_ID and VISIT\_NUMBER. |
| **COLLINTERVAL** | (Longitudinal datasets only) Collection interval is the time frame in which the data for the variable or dataset was collected. |
| **ORDER** | (Optional) The order in which VALUES appear on the variable summary report page. If VALUES of a single variable/column data are integers or decimals, leave blank. If VALUES are encoded values, string, or mixed, define the order. VALUES can be ordered by **Frequency** (highest to lowest frequency of VALUES) or by **List** (user specifies order through placement in VALUES columns). For mixed values within a single variable/column of data, see examples: "age" and "weight" in example file 3b\_dbGaP\_SubjectPhenotypesDD.xlsx. |
| **VALUES** | List of all unique values and/or descriptions of all encoded values, one value per cell. Encoded values are defined as a value and its meaning. For example, if a data file contains a variable named “EDUCATION” and its data values are “1, 2, 3, and 99,” these coded values will need to be defined in the data dictionary. The format of an encoded value is **VALUE=MEANING**. Therefore, in the data dictionary, there should be 4 separate data cells filled out with the following: 1=Completed High School, 2=Completed College, 3=Completed Graduate School, 99=Unknown.  The “VALUES” header must be the last column header (farthest right in the table). It should appear only in the column above the first encoded value that is listed. The remaining column header cells should be left blank. The script will identify the first code meanings and continue right until there are no more code meanings.  For example, if the variable “SEX,” has 3 encoded values: 1=Male, 2=Female and 3=Unknown, the column header “VALUES” will appear only above the cell that contains 1=Male. 1=Male, 2=Female and 3=Unknown will be listed in three separate cells next to each other. The header column cells above “2=Female” and “3=Unknown” should be left blank. |

|  |  |
| --- | --- |
| Action | * Fill in the [column headers](#Column_Headers) listed above: VARNAME, VARDESC, etc. * Under the column headers, fill in each variable found in the matching Phenotype Data File and all variable information that can be provided. Variables must have unique names within a file and preferably within a study. The syntax, case and spacing of the VARNAME must match the VARNAME provided in the matching Phenotype Data File. The DD and data file should have the same number of variables. * Check formatting. For tab-delimited txt files, open in a txt editor, and remove any unintended quotes (“””). Remove empty rows between data values. |
| Accepted Format | * Each DD must be in a separate file. Please do not submit multiple DD worksheets per file and especially please do not submit the DD and data as different worksheets in the same file. * Tab-delimited txt\*, xls and xlsx only. * \*Tab-delimited txt files are preferable. The final dump files provided to Authorized Users of the study will be in the tab-delimited txt format.   (File names should not contain special characters)   * All data file names must be submitted without spaces, hyphens, brackets, periods, or forward (/) or backward slashes (\). |
| Common errors that require resubmission | * Non-ascii characters, new line feeds or carriage return characters (they sometimes may appear like a square or a question mark in a box). * Encoded values for a single variable appear in multiple rows or multiple encoded values appear in a single cell. * There are additional rows above the expected column headers. |

**b.** **Subject Consent Data File - only 1 subject consent file should be submitted to dbGaP per study**

The Subject Consent Data File contains a complete list of non-redundant, de-identified subjects IDs (click here to see definition of [dbGaP Subject](#dbGaP_Subject)), their associated consent groups, and affection statuses. To see an example of a Subject Consent Data File, open **4a\_dbGaP\_SubjectDS.txt** in the directory Phenotype\_Data using Excel. Subjects should include:

1. All subject IDs with submitted phenotype and/or molecular data (ex. genotypes, SRA, GEO) need to be included.
2. Pedigree subject IDs without associated data that were created to provide pedigree relationship information.
3. Subject IDs of genotyping control subjects, such as HapMap subjects. Please work with your genotyping center to identify the Subject IDs that were used as controls.

Additional Subject IDs that are not found in phenotype or molecular data files will be reported to the submitter by dbGaP.

If the data are stored under an alternative de-identified ID, then there should be a mapping of the primary Subject ID to its alias ID in this file.

|  |  |
| --- | --- |
| Action | **VARNAMES**:**Var**iable **names** are listed in the header row. There should be one variable per column. The VARNAME must not contain forward (/) or backward slashes (\) or commas (,).  Required Variables:   * Column 1 – **SUBJECT\_ID**: In a single column, list all the unique, de-identified Subject IDs that will have phenotypes, genotypes, molecular data, pedigree information submitted for your study. Please review the section regarding [Subject De-identification](#Subject_Deidentification). * Column 2 – **CONSENT**: Each Subject ID may be assigned to only one single consent group. Consent groups are determined by your institution and designated Data Access Committee (DAC). List every subject’s Consent group in the column. The consent value must be an integer (1,2,3…). Please use “0” only for subjects used as genotyping controls and/or pedigree linking members (i.e. subject IDs without any submitted phenotype and/or molecular data). * The consent group description should match the consent group name and Data Use Limitation (DUL) provided by your Program Officer (PO) and/or GWAS Program Administrator (GPA). For questions regarding consent name and DUL, please contact your PO or GPA.   Required Variables if applicable (Please do not include the column headers if there are no data entries):   * Columns 3 and 4 – **SUBJECT\_SOURCE** and **SOURCE\_SUBJECT\_ID**: For subjects originating from a shared source (such as a public repository, consortium, institute, study, etc) or for subjects with alias IDs, please include these 2 variables. The data will be marked with error if only one of the variables is included. The **Subject Source (SUBJECT\_SOURCE)** is the name of the third party source, public repository, consortium, institute, or study that corresponds to the subject. The **Source Subject ID (SOURCE\_SUBJECT\_ID)** is the de-identified alias Subject ID used in the public repository, consortium, institute, or study from where the subject has been obtained. The SOURCE\_SUBJECT\_ID maps to the SUBJECT\_ID. If a subject does not have an alias, please leave blank. Please do not fill with N/A or any other string text.   \* For referencing HapMap subjects from Coriell, the SUBJECT\_SOURCE value should be written as “Coriell.” The SOURCE\_SUBJECT\_ID should be written as the de-identified subject ID assigned by Coriell.  \* For referencing dbGaP subjects, the SUBJECT\_SOURCE value should be written as “dbGaP.” The SOURCE\_SUBJECT\_ID should be written as the dbGaP assigned Subject ID.  \* The SUBJECT\_ID and SOURCE\_SUBJECT\_ID can have identical or different IDs.  \* For Subject IDs that map to more than one alias, please contact dbGaP (Example: SUBJECT\_ID 101 is also known as NA1111(Coriell) and 45678(NHGRI)).   * Column 5 – **AFFECTION\_STATUS**: Indicates the presence or absence of disease, or case/control/other status of the subject. If the variable “AFFECTION\_STATUS” applies, please include in column 5 if SUBJECT\_SOURCE and SOURCE\_SUBJECT\_ID are filled in. If there are no SUBJECT\_SOURCE and SOURCE\_SUBJECT\_ID variables, AFFECTION\_STATUS should be included in column 3. If the AFFECTION\_STATUS variable is already included in a Subject Phenotype Data File, it does not need to be included in the Subject Consent File and vice versa. * All other variables containing subject phenotypes should be included in the Subject Phenotype Data File. * Check formatting. For tab-delimited txt files, open in a txt editor, and remove any unintended quotes (“””). Remove empty rows between data values. |
| Accepted Format | * Each data worksheet must be in a separate file. Please do not submit multiple data worksheets per file and especially please do not submit the data and data dictionary as different worksheets in the same file. * Tab-delimited txt\*, xls and xlsx only. * \*Tab-delimited txt files are preferable. The final dump files provided to Authorized Users of the study will be in the tab-delimited txt format.   (File names should not contain special characters)   * All data file names must be submitted without spaces, hyphens, brackets, periods, or forward (/) or backward slashes (\). |
| Common errors that require resubmission | * Multiple Subject Consent Files are submitted for one study instead of combined into a single file. * Subject IDs are not unique and repeated in multiple rows. Sometimes, this is an indication that the Subject IDs do not refer to a subject (click here to see definition of [dbGaP Subject](#dbGaP_Subject)), but to a sample (click here to see definition of [dbGaP Sample](#dbGaP_Sample)). * Subjects who are missing consents or do not have the correct consent value. * Subjects who have the consent value “0,” but are included in a Subject Phenotype Data File or have samples included in the Sample Attributes Data File. A subject must be consented with an integer value in the Subject Consent File in order for the subject to be included in the Subject Phenotype Data File and its sample(s) to be included in the Sample Attributes Data File. * Subjects who are listed in the Subject Consent Data File, but not found in the phenotype data files, subject-sample mapping file, pedigree file, molecular data file and are not a genotyping controls. * SUBJECT\_SOURCE and SOURCE\_SUBJECT\_ID are not submitted in conjunction. * Source Subject IDs are not de-identified. * IDs include characters other than the following approved characters: English letters, Arabic numerals, period (.), hyphen (-), underscore(\_), at symbol (@), and the pound sign (#). |

**Subject Consent Data Dictionary File**

Please see [Phenotype Data Dictionary Files](#Phenotype_DD) for the general format of the data dictionary file. Variable names in the Subject Consent Data Dictionary File must appear identically to those in the Subject Consent Data File. The column headers “VARNAME,” “VARDESC,” and “VALUES” are required; other headers may be included whenever applicable. To see an example of a Subject Consent Data Dictionary File, open **4b\_dbGaP\_SubjectDD.xlsx** in the directory Phenotype\_Data using Excel.

|  |  |
| --- | --- |
| Action | * Fill in the column headers: VARNAME, VARDESC, TYPE, VALUES and any additional column headers that apply. * In the VARNAME column, fill in all variable names that have been used in the Subject Consent Data File. Fill in the remaining columns. * Specifically for the “CONSENT” variable, please code using integers. If the consent group "General Research Use" is present, please use the code value "1" for this group (1=General Research Use (GRU)). If there is a second consent group, please use the same format 2=[Consent Name (abbreviation)] and so forth. **Note**: Please do not fill in "0" as an encoded value in the VALUES column for the “CONSENT” variable. dbGaP will automatically fill in “*0=Subjects did not participate in the study, did not complete a consent document and are included only for the pedigree structure and/or genotype controls, such as HapMap subjects*” as “0” is a reserved value in the dbGaP system. * Check formatting. For tab-delimited txt files, open in a txt editor, and remove any unintended quotes (“””). Remove empty rows between data values. |
| Accepted Format | * Each DD must be in a separate file. Please do not submit multiple DD worksheets per file and especially please do not submit the DD and data as different worksheets in the same file. * Tab-delimited txt\*, xls and xlsx only. * \*Tab-delimited txt files are preferable. The final dump files provided to Authorized Users of the study will be in the tab-delimited txt format.   (File names should not contain special characters)   * All data file names must be submitted without spaces, hyphens, brackets, periods, or forward (/) or backward slashes (\). |
| Common errors that require resubmission | * The VALUES for the “CONSENT” variable are not provided. * The consent groups listed for the “CONSENT” variable does not match the consent name and Data Use Limitation provided by your study’s Program Officer (PO) and/or GWAS Program Administrator (GPA). * SUBJECT\_SOURCE and SOURCE\_SUBJECT\_ID are not submitted together. |

**c.** **Subject Sample Mapping Data File - only 1 subject sample mapping file should be submitted to dbGaP per study**

The Subject Sample Mapping Data File relates each Subject ID (click here to see definition of [dbGaP Subject](#dbGaP_Subject)) for which phenotype data has been submitted to each genotyped/sequenced Sample ID (click here to see definition of [dbGaP Sample](#dbGaP_Sample)). To see an example of a Subject Sample Mapping Data File, open **5a\_dbGaP\_SubjectSampleMappingDS.txt** in the directory Phenotype\_Data using Excel.

If the primary Sample IDs have alias IDs, the mapping of the primary Sample ID to its alias ID should also be included in this file.

|  |  |
| --- | --- |
| Action | **VARNAMES**:**Var**iable **names** are listed in the header row. There should be one variable per column. The VARNAME must not contain forward (/) or backward slashes (\) or commas (,).  Required Variables:   * Columns 1 and 2 – **SUBJECT\_ID** and **SAMPLE\_ID**: List the mapping of de-identified Subject IDs to all unique de-identified Sample IDs that have genotype/sequence submitted to the dbGaP. These also include genotyping controls (e.g., HapMap subjects). Please work with your genotyping center to identify the Sample IDs that were used as genotyping controls. A subject with multiple samples should be listed in multiple rows with the same Subject ID, but distinct Sample IDs. Similarly, if one patient (Subject ID) gave one sample, but that sample was processed differently to generate 2 genotyping/ sequencing runs, there would be two rows in the Subject Sample Mapping Data File, both using the same Subject ID, but having different Sample IDs. * Column 3 – **SAMPLE\_USE**: Standardized terms describing the use of the sample. Please use the SAMPLE\_USE values listed in the first column: <http://www.ncbi.nlm.nih.gov/projects/gap/submission/GetSampleUseTypes.cgi>. For a sample ID with multiple uses, please delimit the values with a semicolon (ex. Seq\_DNA\_WholeExome; Seq\_DNA\_SNP).   Required variables if applicable (Please do not include the column header if all values are empty):   * Columns 3 and 4 – **SAMPLE\_SOURCE** and **SOURCE\_SAMPLE\_ID**: For samples originating from a shared source (such as a public repository, consortium, institute, study, etc) or for samples with alias IDs, please include these 2 variables. The data will be marked with error if only one of the columns is submitted. The **Sample Source (SAMPLE\_SOURCE)** is the name of the third party source, public repository, consortium, institute, or study from where the sample has been obtained. The **Source Sample ID (SOURCE\_SAMPLE\_ID)** is the de-identified alias Sample ID used in the public repository, consortium, institute, or study. The SOURCE\_SAMPLE\_ID maps to the SAMPLE\_ID. If a sample does not have an alias, please leave blank. Please do not fill with N/A or any other string text.   \* For referencing HapMap samples from Coriell, the SAMPLE\_SOURCE value should be written as “Coriell.” The SOURCE\_SAMPLE\_ID should be written as the de-identified Sample ID assigned by Coriell.  \* For referencing dbGaP samples, the SAMPLE\_SOURCE value should be written as “dbGaP.” The SOURCE\_SAMPLE\_ID should be written as the dbGaP assigned Sample ID.  \* The SAMPLE\_ID and SOURCE\_SAMPLE\_ID can have identical or different IDs.  \* For Sample IDs that map to more than one alias, please contact dbGaP (Example: SAMPLE\_ID 201 is also known as NA2111(Coriell) and 25678(NHGRI)).  Optional variables:   * Any additional plate/well information which might be useful in resolving potential sample mix-ups can be included as a variable. All other variables containing sample attributes should be included in the Sample Attributes Data File. * Check formatting. For tab-delimited txt files, open in a txt editor, and remove any unintended quotes (“””). Remove empty rows between data values. |
| Accepted Format | * Each data worksheet must be in a separate file. Please do not submit multiple data worksheets per file and especially please do not submit the data and data dictionary as different worksheets in the same file. * Tab-delimited txt\*, xls and xlsx only. * \*Tab-delimited txt files are preferable. The final dump files provided to Authorized Users of the study will be in the tab-delimited txt format.   (File names should not contain special characters)   * All data file names must be submitted without spaces, hyphens, brackets, periods, or forward (/) or backward slashes (\). |
| Common errors that require resubmission | * Multiple Subject Sample Mapping Files are submitted for one study instead of combined into a single file. * Sample IDs are not unique and repeated in multiple rows. Sometimes, this is an indication that the Sample IDs do not refer to a sample (click here to see definition of [dbGaP Sample](#dbGaP_Sample)), but to a subject (click here to see definition of [dbGaP Subject](#dbGaP_Subject)). * A single Sample ID maps to two different Subject IDs. * Samples map to subjects who are missing consents or do not have the correct consent value listed in the Subject Consent Data File. * SAMPLE\_SOURCE and SOURCE\_SAMPLE\_ID are not submitted together. * IDs include characters other than the following approved characters: English letters, Arabic numerals, period (.), hyphen (-), underscore(\_), at symbol (@), and the pound sign (#). |

**Subject Sample Mapping Data Dictionary File**

Please see [Phenotype Data Dictionary Files](#Phenotype_DD) for the general format of the data dictionary file. Variable names in the Subject Sample Mapping Data Dictionary File must appear identically to those in the Subject Sample Mapping Data File. The column headers “VARNAME” and “VARDESC” are required; other headers may be included whenever applicable. To see an example of a Subject Sample Mapping Data Dictionary File, open **5b\_dbGaP\_SubjectSampleMappingDD.txt** in the directory Phenotype\_Data using Excel.

|  |  |
| --- | --- |
| Action | * Fill in the column headers: VARNAME, VARDESC and any additional column headers that apply. * For SAMPLE\_USE variable, please leave the encoded values blank. dbGaP will automatically fill in the coded meanings using the descriptions found here: <http://www.ncbi.nlm.nih.gov/projects/gap/submission/GetSampleUseTypes.cgi>. * Fill in all variable names that have been used in the Subject Sample Mapping Data File. Fill in the remaining columns. * Check formatting. For tab-delimited txt files, open in a txt editor, and remove any unintended quotes (“””). Remove empty rows between data values. |
| Accepted Format | * Each DD must be in a separate file. Please do not submit multiple DD worksheets per file and especially please do not submit the DD and data as different worksheets in the same file. * Tab-delimited txt\*, xls and xlsx only. * \*Tab-delimited txt files are preferable. The final dump files provided to Authorized Users of the study will be in the tab-delimited txt format.   (File names should not contain special characters)   * All data file names must be submitted without spaces, hyphens, brackets, periods, or forward (/) or backward slashes (\). |
| Common errors that require resubmission | * SAMPLE\_SOURCE and SOURCE\_SAMPLE\_ID are not submitted in conjunction. |

**d.** **Pedigree Data File - only 1 pedigree file should be submitted to dbGaP per study if applicable**

The Pedigree Data File lists the genealogical relationships of subjects within a study. To see an example of a Pedigree Data File, open **6a\_dbGaP\_PedigreeDS.txt** in the directory Phenotype\_Data using Excel.

|  |  |
| --- | --- |
| Action | **VARNAMES**: **Var**iable **names** are listed in the header row. There should be one variable per column. The VARNAME must not contain forward (/) or backward slashes (\) or commas (,).  Required variables:   * Column 1 – **FAMILY\_ID:** List de-identified Family IDs. The Family ID is also referred to as the Pedigree ID. * Column 2 – **SUBJECT\_ID:** List de-identified Subject IDs. Subject IDs should also include IDs of each unique father and mother. These are the same primary Subject IDs that are listed in the Subject Consent Data File. * Columns 3 and 4 – **FATHER** and **MOTHER**: Every individual father must have a unique, de-identified Father ID; every individual mother must have a unique, de-identified Mother ID. The Father ID and Mother ID should not be identical. Use 0 (zero) or leave blank for founders or marry-ins (parents not specified) in a pedigree. Each unique Father ID and unique Mother ID should also be listed in the SUBJECT\_ID column of both the Pedigree Data File and the Subject Consent Data File. * Column 5 – **SEX**: List the gender of the subject in the SUBJECT\_ID column.   Required variables if applicable (Please do not include the column header if all values are empty):   * **TWIN\_ID**: Identify monozygotic twins or multiples of the same family with Twin IDs. Twins or multiples of the same family should share the same TWIN\_ID, but be assigned different SUBJECT\_IDs. Zygote pairs/multiples should be numbered sequentially. If twins are dizygotic, the TWIN\_ID could be used, but an additional variable should be included to specify that the twins are dizygotic. * Check formatting. For tab-delimited txt files, open in a txt editor, and remove any unintended quotes (“””). Remove empty rows between data values. |
| Accepted Format | * Each data worksheet must be in a separate file. Please do not submit multiple data worksheets per file and especially please do not submit the data and data dictionary as different worksheets in the same file. * Tab-delimited txt\*, xls and xlsx only. * \*Tab-delimited txt files are preferable. The final dump files provided to Authorized Users of the study will be in the tab-delimited txt format.   (File names should not contain special characters)   * All data file names must be submitted without spaces, hyphens, brackets, periods, or forward (/) or backward slashes (\). |
| Common errors that require resubmission | * Father and Mother IDs are not listed in the SUBJECT\_ID column. * Father and Mother IDs are not unique, e.g., Father IDs are all 1’s and Mother IDs are all 2’s. One option is to append Family IDs to make them unique, e.g., Father ID = 103\_1 and Mother ID = 103\_2. * SEX value in the Pedigree Data File conflicts with either 1) the subject being a Father or Mother or 2) other Phenotype Data Files. * Subjects who are founders or marry-ins are not listed as having a Father and Mother ID of blank or “0.” |

**Pedigree Data Dictionary File**

Please see [Phenotype Data Dictionary Files](#Phenotype_DD) for the general format of the data dictionary file. Variable names in the Pedigree Data Dictionary File must appear identically to those in the Pedigree Data File. The column headers “VARNAME,” “VARDESC,” and “VALUES” are required; other headers may be included whenever applicable. To see an example of a Pedigree Data Dictionary File, open **6b\_dbGaP\_PedigreeDD.xlsx** in the directory Phenotype\_Data using Excel.

|  |  |
| --- | --- |
| Action | * Fill in the column headers: VARNAME, VARDESC, TYPE, VALUES and any additional column headers that apply. * Fill in all variable names that have been used in the Pedigree Data File. Fill in the remaining columns. * For the encoded VALUES of the variable “SEX,” please use “1=male” or “M=male” and “2=female” or “F=female.” * Check formatting. For tab-delimited txt files, open in a txt editor, and remove any unintended quotes (“””). Remove empty rows between data values. |
| Accepted Format | * Each DD must be in a separate file. Please do not submit multiple DD worksheets per file and especially please do not submit the DD and data as different worksheets in the same file. * Tab-delimited txt\*, xls and xlsx only. * \*Tab-delimited txt files are preferable. The final dump files provided to Authorized Users of the study will be in the tab-delimited txt format.   (File names should not contain special characters)   * All data file names must be submitted without spaces, hyphens, brackets, periods, or forward (/) or backward slashes (\). |
| Common errors that require resubmission | * Variable “SEX” encoded values do not correspond to subject being a mother or father. |

1. **Study Documents**
2. Describe the study’s intent, design and background
3. Represent/describe the source of the data (e.g., questionnaires, interview forms, protocols)
4. Describe data collection and analysis methods, including:
   1. Which question(s)/section(s) of the documents are associated with each variable (please be specific).
   2. Methods (e.g. derivations/algorithms) that were used to obtain the final data values especially if data values of one variable were derived from more than one source/question/ measurement or variable.

|  |  |
| --- | --- |
| Action | * Submit relevant consent forms, protocols, questionnaires, Manual of Procedures, Manual of Operations, forms, etc. |
| Editorial Policy | It is assumed that the documents submitted to dbGaP have been proofread for content. dbGaP will not perform any editorial work to the documents (beyond fixing typos) unless a new version of the document is submitted that includes content changes. Documents that contain subject information in violation of HIPAA rules will not be processed; they need to be resubmitted with appropriate redactions. |
| Previously published work | If you submit a published work (article, review, book chapter, etc.) or portion of a published work as a dbGaP document, please include documentation that authorizes it to be freely distributed on the dbGaP site. If you are unsure about the copyright status of a document, please contact the publisher or owner of the work.  NIH does not claim copyright on any submitted documents, but it must be given nonexclusive rights to freely distribute all documents submitted to dbGaP.  The submitting PI is responsible for ensuring that applicable copyright or other restrictions are clearly indicated on all documents. The PI is also responsible for obtaining the necessary permissions to make such documents publicly available on the dbGaP site, and for conveying these permissions to the dbGaP project. NIH reserves the right to withhold or withdraw a copyrighted document from the dbGaP site if the necessary permissions are not included with the document. Neither dbGaP nor NIH will contact others to obtain such permissions. |
| Accepted Format | docx, doc, xls, xlsx, txt, and pdf (Other formats should be discussed with dbGaP prior to submission, e.g., xml, html, etc.). |
| Common errors that require resubmission | * Documents in violation of HIPAA rules (See “Editorial Policy”). * Documents are not in English. |

1. **Molecular Data**

Molecular Data includes the following if applicable:

1. Genotypes
2. High Throughput Sequencing submitted to [NCBI Sequence Read Archive](http://www.ncbi.nlm.nih.gov/sra) (SRA) and distributed via dbGaP Authorized Access in collaboration with SRA
3. VCF or other matrices containing SNPs, CNVs, and/or structural variants called from sequence data
4. Copy Number Variation (CNV) is coordinated with [NCBI dbVar](http://www.ncbi.nlm.nih.gov/dbvar)
5. Sample Trace Mapping if trace data were submitted to the [NCBI Trace Archive](http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?)
6. Sample GEO Mapping if microarray, next-generation sequencing, and other forms of high-throughput functional genomic data were submitted to [NCBI GEO](http://www.ncbi.nlm.nih.gov/geo/)
7. Sample GenBank Mapping if genetic sequences were submitted to [NCBI GenBank](http://www.ncbi.nlm.nih.gov/genbank/)

|  |  |
| --- | --- |
| Action | * If **Genotypes** are applicable: Open **Genoptype\_Submission\_Guide.docx** in the directory Molecular\_Genetic\_Data and follow the instructions in the guide. * For sequences being submitted to **SRA**, [Subject Consent Data File](#Subject_Consent_Data_File) and [Subject Sample Mapping Data File](#Subject_Sample_Mapping_Data_File) must be submitted to dbGaP first. DbGaP will load the subject and sample IDs and consents, and make them available to [BioSample](http://www.ncbi.nlm.nih.gov/biosample)/[SRA](http://www.ncbi.nlm.nih.gov/sra). Once loaded, we will provide you with an SRA contact person to submit sequences. Please see this handbook on the SRA submitting process: <http://www.ncbi.nlm.nih.gov/books/NBK242619/>. Please let SRA know if you will be generating XML or will need a spreadsheet. * For **Variant Call Format** **(vcf)** files being submitted, please refer to p.4 #3 of the “Genotype\_Submission\_Guide.docx” under the directory Molecular\_Genetic\_Data for instructions. * **Copy Number Variation** data is coordinated with the NCBI's dbVar. Individual level data will be configured and released via controlled access. Summary level (probe/primer and other assay and frequency information) copy number variation data will be released by the public dbVAR. Please see the dbVAR submission guide if your study includes CNV data: <http://www.ncbi.nlm.nih.gov/dbvar/content/submission/>. * If **Sample Trace Mapping** is applicable: Please see **dbGaP\_SampleTraceMappingDS.txt** and **dbGaP\_SampleTraceMappingDD.xlsx** in the directory Molecular\_Genetic\_Data using Excel.   **VARNAMES**:**Var**iable **names** are listed in the header row. There should be one variable per column. The VARNAME must not contain forward (/) or backward slashes (\) or commas (,).  Required Variables:  Columns 1 and 2 – **SAMPLE\_ID** and **TRACEID**: List the mapping of de-identified Sample IDs to all unique Trace IDs that have trace data submitted to [NCBI Trace Archive](http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?). A sample with multiple traces should be listed in multiple rows with the same Sample ID, but distinct Trace IDs.  \*Please also submit in a separate file the matching Sample Trace Mapping Data Dictionary.   * If **Sample GEO Mapping** is applicable: Please see **dbGaP\_SampleGEOMappingDS.txt** and **dbGaP\_SampleGEOMappingDD.xlsx** in the directory Molecular\_Genetic\_Data using Excel.   **VARNAMES**:**Var**iable **names** are listed in the header row. There should be one variable per column. The VARNAME must not contain forward (/) or backward slashes (\) or commas (,).  Required Variables:  Columns 1 and 2 – **SAMPLE\_ID** and **GEO\_ACCESSION**: List the mapping of de-identified Sample IDs to all unique GEO accessions that have data submitted to [NCBI GEO](http://www.ncbi.nlm.nih.gov/geo/). A sample mapped to multiple GEO accessions should be listed in multiple rows with the same Sample ID, but distinct GEO IDs.  \*Please also submit in a separate file the matching Sample GEO Mapping Data Dictionary.   * If **Sample GenBank Mapping** is applicable: Please see **dbGaP\_SampleGenBankMappingDS.txt** and **dbGaP\_SampleGenBankMappingDD.xlsx** in the directory Molecular\_Genetic\_Data using Excel.   **VARNAMES**:**Var**iable **names** are listed in the header row. There should be one variable per column. The VARNAME must not contain forward (/) or backward slashes (\) or commas (,).  Required Variables:  Columns 1 and 2 – **SAMPLE\_ID** and **GENBANK\_ACCESSION**: List the mapping of de-identified Sample IDs to all unique GenBank accessions that have data submitted to [NCBI GenBank](http://www.ncbi.nlm.nih.gov/genbank/). A sample mapped to multiple Genbank accessions should be listed in multiple rows with the same Sample ID, but distinct GEO IDs.  \*Please also submit in a separate file the matching Sample GenBank Mapping Data Dictionary. |

1. **Association Analysis Data**

Analysis of association of genotype and phenotypic trait(s)

|  |  |
| --- | --- |
| Action | * For **Association** **Analyses** **Data** being submitted: See **Association\_Analysis\_Submission\_Guide.docx** in the directory **Association\_Analysis\_Data** and follow the instructions in d**bGaP\_Association\_Analysis.xlsx** for Case-Controls (Worksheet 1) or Others (Worksheet 2). If GWAS results are submitted as outputs of the software, please give brief descriptions of the headers, indicating the linking-columns and/or relationships when several files are involved. |

Submitters can contact dbGaP at: [dbgap-help@ncbi.nlm.nih.gov](mailto:dbgap-help@nvbi.nlm.nih.gov).

**dbGaP Contacts:**

|  |  |
| --- | --- |
| **Michael Feolo**  Staff Scientist  National Center for Biotechnology Information  45 Center Drive, Bldg.45, Room 4AN.12B  Bethesda, MD  20892  301.402.2874  [feolo@ncbi.nlm.nih.gov](mailto:feolo@ncbi.nlm.nih.gov) | **Lora Ziyabari**  Contractor (MSC)  National Center for Biotechnology Information  45 Center Drive, Bldg.45, Room 4AN.12D-59  Bethesda, MD  20892  301.443.2455  [ziyabarl@ncbi.nlm.nih.gov](mailto:ziyabarl@ncbi.nlm.nih.gov) |