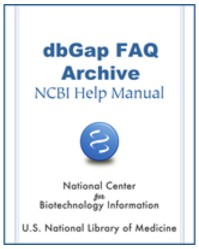
NLM Citation: GaP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2009-. Data Sample and Subject ID Mapping.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Data Sample and Subject ID Mapping

Sample and Subject ID Mapping of Pheno, Genotype, and Sequencing Data

How can I map subject and sample IDs found in phenotype, genotype, and sequencing data files?

The dbGaP phenotype, genotype, and sequencing data (including BAM, SRA data etc.) are often submitted and processed separately. One of the consequences of it is that the header names of IDs used in different data files may be in different naming formats. The following information may help you to get IDs mapped cross all data files.

1. Phenotype subject, sample ID mapping

   The master mapping files between subject and sample IDs can be found from the files that have the phrase "**_Subject**", or "**_Sample**" or "**_Pedigree**" embedded in the file name. For example:

   phs000094.v1.pht001136.v1.p1.Oral_Clefts_Subject.MULTI.txt

   phs000094.v1.pht001138.v1.p1.Oral_Clefts_Sample.MULTI.txt

   phs000094.v1.pht001137.v1.p1.Oral_Clefts_Pedigree.MULTI.txt

   In the authorized access system, these files are placed together with phenotype files in the file selection tree. The file selection tree can be found in the "Access Request" page under "My Request" tab.

2. Genotype ID mapping

   The sample and subject ID mapping information of genotype files can be found in a file packed in the tarball that has the phrase "**sample-info**" embedded in the taball name. For example:

   phg000054.v1.p1.GENEVA_OralClefts.sample-info.MULTI.tar

   Please note that the **header title of IDs in the sample-info file may not be exactly identical to those used in the master mapping files** mentioned above. The corresponding IDs in the master mapping file should identified easily based the face meaning of ID headers in the genotype sample-info file.

3. SRA sample ID mapping

   The SRA samples are given independent IDs at the different stage of data processing, handling, and archiving for different purposes. For example most of the SRA samples distributed through the dbGaP have submitted_sample_id, sra_accession, sra_sample_id, and dbgap_sample_id. The mapping information of these IDs can be found in a manifest file available on the "Access Request" page. The following is how to locate the manifest file:

   Login to the dbGaP account, go to "My Request" tab, find the data access request of interest from the request list, and click on the "Request Files" link in the "Actions" column. A manifest that contains SRA sample ID mapping is available through a link named "Dataset Manifest".

(15/15/2014)

The Description of Sample, Subject IDs Used in dbGaP Data Files

I see so many IDs in dbGaP files and wonder what exactly they are. Could you provide more information about them?

Answer:

1. **dbGaP SampID**

   The dbGaP Sample ID is a dbGaP assigned accession to the submitted SAMPID. Please see SAMPID for more information. The dbGaP SampID is included as a column in the final phenotype dump files whenever there is a submitted sample ID column.

2. **dbGaP SampID**

   The dbGaP Sample ID is a dbGaP assigned accession to the submitted SAMPID. Please see SAMPID for more information. The dbGaP SampID is included as a column in the final phenotype dump files whenever there is a submitted sample ID column.

3. **dbGaP SubjID**

   The dbGaP Subject ID is a dbGaP assigned accession to the submitted SUBJID. Please see SUBJID for more information. The dbGaP SubjID is included as a column in the final phenotype dump files whenever there is a submitted subject ID column.

   The dbGaP Subject ID is unique cross all dbGaP studies, which means that if a subject is known to have participated in multiple studies that have been submitted to dbGaP, the same dbGaP SubjID will be assigned to the individual across multiple studies, though the submitted subject ID may be different.

4. **SUBJID**:

   The SUBJID is submitted subject ID and is included in the Subject Consent Data File, Subject Sample Mapping Data File, Pedigree Data File (if available), and all Subject Phenotype Data Files. A dbGaP Subject is defined as a single human person/individual/patient that arises from a single germline. Each subject has been assigned a single, unique, de-identified Subject ID. Subject IDs should be an integer or string value. Only the following characters can be included in the ID: English letters, Arabic numerals, period (.), hyphen (-), underscore (_), at symbol (@), and the pound sign (#). In addition to the submitted Subject ID, dbGaP will assign a dbGaP Subject ID that will be included in the final phenotype dump files along with the submitted Subject ID.

5. **SAMPID**

   The SAMPID is the submitted sample ID and is included in the Subject Sample Mapping Data File and Sample Attributes Data File. A dbGaP Sample is defined as 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 BioSamples. A single subject can have multiple samples, but a single sample cannot 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. Only the following characters can be included in the ID: English letters, Arabic numerals, period (.), hyphen (-), underscore (_), at symbol (@), and the pound sign (#). In addition to the submitted Sample ID, dbGaP will assign a dbGaP Sample ID that will be included in the final phenotype dump files along with the submitted Sample ID. For example, if one patient (subject ID) gave one sample, and that sample was processed differently to generate two sequencing runs or one sequencing run and 1 genotyping array, there would be two rows, both using the same subject ID, but having 2 unique sample IDs. The SAMPIDs listed in the Subject Sample Mapping Data File should be identical to the samples found in the genotype and SRA Data.

6. **SOURCE_SUBJID and SUBJ_SOURCE**

   For subjects originating from a shared source (such as a public repository, consortium, institute, study, etc.) or for subjects with alias IDs, these 2 variables will be included in the Subject Consent Data File. The **Subject Source (SUBJ_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_SUBJID)** 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_SUBJID maps to the SUBJID.

   For referencing HapMap subjects from Coriell, the SUBJ_SOURCE value is written as "Coriell." The SOURCE_SUBJID should be written as the de-identified subject ID assigned by Coriell.

7. **SEX**

   The gender variable can be included in a subject phenotype data file or in a pedigree file if a pedigree file is available.

8. **FAMID**

   The family ID is found in the pedigree file if a pedigree file is available. FAMID is a column of de-identified Family IDs. The Family ID is also referred to as the Pedigree ID. The family ID should be the same for individuals belonging in the same biological family.

9. **FATHER and MOTHER**

   Every individual father has a unique, de-identified Father ID; every individual mother has a unique, de-identified Mother ID. The Father ID and Mother ID are not identical. 0 (zero) or blank is filled in for founders or marry-ins (parents not specified) in a pedigree. Each unique Father ID and unique Mother ID is also listed in the SUBJID column of both the Pedigree Data File and the Subject Consent Data File.

10. **CONSENT**

    Every subject 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 that appears in a Sample Attribute Data File must belong to a consented subject. The consent information is listed in the Subject Consent Data File. Each subject can only belong to a single consent group. The consents are determined by the submitter, their IRB, their GPA (Genome Program Administrator) along with the DAC (Data Access Committee). All data is parsed into its respective consent groups for download.

(10/25/2012)