# GaP FAQ Archive

Last Updated: April 4, 2014

National Center for Biotechnology Information (US)
Bethesda (MD)

National Center for Biotechnology Information (US), Bethesda (MD)

NLM Citation: GaP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2009-.

This book contains actual questions asked by dbGaP users. We have tried to compile questions and answers that are representative of the entire spectrum of questions that our users ask, regardless of whether or not they are frequently asked. It is therefore likely that the question you have now, or a variant of it, may have been asked before, and may be in this book. The book will be updated with new questions regularly.

The questions and their answers have been edited for clarity and to protect the privacy of our users.

# Table of Contents

# Data Content

Created: October 21, 2008; Updated: March 23, 2009.

This section of the dbGaP FAQ Archive contains general information about the nature of the data contained in dbGaP: information about the characteristics of a specific study's data; general information that applies to all data sets; as well as whether or not certain types of data or tools are part of dbGaP's content.

**To begin searching this section of the dbGaP FAQ Archive, you can either**:

- Enter your search word(s) text in the text box at the top of the page and click on the "Go" button,

  OR

- Click on any of the "Data Content" sub-categories listed in the navigation box on the right side of the page to navigate to the sub-category of your choice.

# Individual-level Data: General Questions

Created: October 21, 2008; Updated: March 23, 2009.

## Data Available through the dbGaP

**Can you tell me what of kind of data hosted in the dbGaP?**

The dbGaP archives and distributes the results of studies that have investigated the interaction of genotype and phenotype. Such studies include genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits. The individual level data hosted at the dbGaP is distributed through a controlled access system. The types of data distributed through the dbGaP include phenotype data, association (GWAS) data, summary level analysis data, SRA (Short Read Archive) data, reference alignment (BAM) data, VCF (Variant Call Format) data, expression data, imputed genotype data, image data, etc. (**01/17/13**)

## De-identification of Data

**Can you tell me if specific individuals can be identified in the controlled access data?**

The individual-level data submitted to the dbGaP is required to be de-identified. No names or identifiable information is attached to the data. The genetic fingerprint however is embedded in individual's genotype data, which is not de-identifiable. That is why, to protect individual's privacy, all individual level data is only distributed through the Authorized Access System.(**04/17/2013**)

## Available Data File Format

**What are the formats of phenotype and genotype data files?**

The phenotype tables are rectangular, and in general are constructed where a single row represents each study participant, and each column is a measured trait.

Genotypes are available in several different formats:

- The Matrix format. This is like the phenotype format listed above, except that the rows represent SNPs and the columns represent samples.
- The PLINK format.
- The VCF format.
- An individual format where there is one file for each sample and all the genotypes are listed.

(**07/03/2012**)

## TCGA Data Availability

**Are TCGA data still distributed through the dbGaP?**

Starting from June, 2016, all TCGA data, including the phenotype and sequencing data, are hosted at the Genomics Data Commons website (https://gdc.cancer.gov/).

The dbGaP continues to manage the controlled access approval process through the Authorized Access System.. The TCGA data access request should be made through the dbGaP system in the same way as other dbGaP studies (look for study phs000178). After the request is approved, the approval information will be passed to the Genomic Data Commons system within 24 hours.

The Genomics Data Common website is operated completely independent of the dbGaP. All issues related to that system, such as system login and data download, should be addressed directly to their help-desk.

(**09/21/2017**)

# Analytical Tool Availability

Created: October 21, 2008; Updated: March 23, 2009.

## Tool for analyzing individual level data remotely

**Does dbGaP have online tools I can use to analyze individual-level data remotely?**

The dbGaP does not have any online tools for remote data analysis against individual level data.

(**10/08/08**)

## Tool for analyzing summary level data

**Does dbGaP have online tools that can help me to identify phenotype and SNP associations of the specific gene or chromosomal region related to my research?**

Answer:

1. There is a tool named Phenotype-Genotype Integrator (PheGenI) may be useful to you. Here is a tutorial video about it.

   PheGenI merges NHGRI genome-wide association study (GWAS) catalog with several databases housed at the NCBI. With the tool, user can search SNP genes association results based on chromosomal location, gene, SNP or phenotype. The PheGenI search however limits to the GWAS summary level analysis data hosted in the dbGaP. It is not a data analytical tool directly against individual level data distributed through the dbGaP Authorized Access System. The tool can be accessed from the dbGaP home page through a link named “Phenotype-Genotype Integrator”

   Please note that the PheGenI displays the best p-values from each dbGaP hosted analysis. Many p-values are excluded (basically, those that dbGaP has deemed as not statistically significant.) If you'd like access to all the p-values (both good and bad), then there are a couple of alternatives:

   1. If all you care about are just p-values (and not, say, direction or minor allele frequency), then you can either
      a. examine an analysis using the genome browser: for example from here, and click on "View association results in Genome Browser"
      b. download the p-values in tabular form from the dbGaP public ftp site: for example from here. Click on "Connect to public download site", then "Analyses", then the zip file.
   2. If you require more information than the above alternatives supply, you will have to make data access request through the Authorized Access System.

2. Another tool named Association Results Browser allows the search against the GWAS catalog and summary level analysis data available at the dbGaP.

(**04/17/2013**)

# Data Access Requests

Created: October 21, 2008; Updated: December 11, 2013.

This section of the database of Genotypes and Phenotypes (dbGaP) FAQ Archive contains general information about completing and submitting a Data Access Request (DAR) that, once approved, will allow you authorized access to dbGaP's controlled-access data. The information in this section ranges from password management, to a quick step-by step of how to complete the DAR and electronically sign it. This section also provides instructions about altering a completed and signed DAR, and information about password and request management for Signing Officials (SOs).

**To begin searching this section of the dbGaP FAQ Archive, you can either**:

- Enter your search word(s) text in the text box at the top of the page and click on the "Go" button,

**OR**

- Click on any of the "Data Access Requests" sub-categories listed in the navigation box on the right side of the page to navigate to the sub-category of your choice.

# Starting Point to Applying for dbGaP Data

Created: October 21, 2008; Updated: December 11, 2013.

## Demo Videos and Overview

**Could you give a demo about the process of applying for dbGaP individual level data?**

Several videos currently available on the YouTube may help you with the data access application, project renewal and closeout processes. The titles and links of the videos are dbGaP Data Access Application, Application Renewal, and Project Closeout. Please also see here for a brief overview of the process. (**12/11/2013**)

## Access Individual Level Data

**I am not able to find controlled-access data on dbGaP public web and FTP sites. Where and how can I get them?**

The dbGaP controlled-access data are available only through the dbGaP Authorized Access System. For NIH researchers (intramural researchers), in order to obtain access to genomic data in dbGaP that is available through controlled-access, eligible NIH Institute and Center (IC) intramural scientists and IC extramural program scientific staff must first obtain the approval of their IC. Please see here for more details. For non-NIH researchers (extramural researchers), the Principal Investigator (PI) must be a tenure-track professor, senior scientist, or equivalent, to be able to submit a data access request (DAR) and have a valid eRA Commons account for logging in to the dbGaP system. Please see here for more about how to setup a new eRA Commons account or how to make changes to an existing eRA Commons account.

The dbGaP data access request procedures are summarized in this document. Once you have your account ready, please see here for more about how to make data request. If you have any further questions, please contact the dbgap-help at dbgap-help@ncbi.nlm.nih.gov.

(**07/27/2017**)

## The dbGaP and eRA Account

**What is the relationship between my dbGaP account and eRA account? How to setup a new eRA account and make changes to an existing eRA account?**

The dbGaP Authorized Access System authenticates non-NIH users using the information registered in the NIH eRA Commons. Once your eRA account is setup, you are ready to login to the dbGaP system using the eRA Commons account login credentials. If you have just setup a new eRA Commons account or made changes to an existing eRA Commons account**, please allow one or two days for the new account information to be propagated from the eRA to the dbGaP system**.

With your dbGaP account, you can create dbGaP projects, select datasets of your interest, complete application forms, and finally submit your data requests for approval first by the Signing Official (SO) of your institution and then by the Data Access Committee (DAC) at NIH. Once the requests are approved, you can download the data through your dbGaP account.

If it is the first time for your institution to setup an eRA Commons account, please see here for the information related to institutional eRA account.

The eRA Commons account is not limited to US researchers. Many foreign organizations have already registered with eRA. You may want to first check with your institution to see if there is an institutional standing account. If there is one, the eRA Commons account administrator of your institution should be able to help you.

The eRA system is not operated by the National Center for Biotechnology Information (NCBI) that oversees dbGaP. To get a new eRA Commons account or make any changes to an existing eRA Commons account, such as resetting the login password or changing the email address, please visit the eRA website or directly contact the eRA help desk.

The eRA Commons How To and FAQ sites are often found to be useful.

(**07/27/2017**)

## Institutional eRA Account

**I don't think my institution has an eRA Commons account. Do they need to setup one before I can access dbGaP?**

Yes, your institution will need to register at NIH eRA Commons before you can apply for an eRA Commons accountthat is necessary to access dbGaP. Institutions are usually registered by someone in the business office, since it requires financial information (such as the DUNS number). Once registration is complete, the eRA Commons account administrator of the institution should be able to assist individual investigators under the institution to setup their own eRA Commons account.

account.

(**07/03/2017**)

## DUNS Number

**What is a DUNS number? Do I need a DUNS number for my dbGaP data access request?**

DUNS stands for "Data Universal Numbering System". It is a unique nine-digit numbering system that is used to identify a business. The DUNS numbers are assigned by Dun and Bradstreet and are a part of an institution's account in the NIH eRA electronic record system. The DUNS number is not used in the data access requests generated by dbGaP, although it is required for establishing an institution's eRA Commons account. Please refer to D&B online DUNS request service for more details. The DUNS Number Search page may also be useful.

(**06/29/2017**)

## Principal Investigator (PI) Role Is Required

**I am an Assistant Professor and a Principal Investigator (PI). When trying to login to the dbGaP system, it tells me that I don't have data request privileges for data access. What do I need to proceed?**

In order to access the dbGaP data, you need to be registered as a Principal Investigator (PI) by your institution with your eRA account. There may be three situations in which you are a PI in your institution, but have no privilege to make data request in the dbGaP system.

1. You have just created a new eRA Commons account or the role of the account is just changed to PI. In this case, please allow one to two days for the updated information to be propagated from the eRA to the dbGaP system before trying to login to the dbGaP Authorized Access System.
2. Your eRA Commons account is not new and the role with the account is not newly updated. In this case, we would like you to login to your eRA Commons account directly from eRA Commons to double check the role associated with the account. If the PI role is confirmed, it may suggest problems with the dbGaP system. Please contact dbgap-help at dbgap-help@ncbi.nlm.nih.gov .
3. You have confirmed that your eRA account has no PI role. In this case, you would need to contact the eRA Commons account administrator of your institution

for help. The administrator can change the role of your eRA Commons account. Once the role is changed, please allow one to two days for the update to become effective in the dbGaP system.

(**07/27/2017**)

## For NIH Staff

**I am an investigator at the NIH. What do I need to be able to access the dbGaP controlled-access data?**

In order to obtain access to genomic data in dbGaP that is available through controlled-access, eligible NIH Institute and Center (IC) intramural scientists and IC extramural program scientific staff must first obtain the approval of their IC. After completing the form and obtaining the necessary signatures, please scan the form and email the pdf document to the attention of the Genomic Data Sharing Policy Staff at GDS@mail.nih.gov. Once your completed form is submitted, you will be registered in the dbGaP system as an approved user and will be notified by email when you can proceed to submit data access requests to dbGaP via its Authorized Access System. More instructions about how to access dbGaP data can be found here. If you have any further questions, please contact the dbgap-help at dbgap-help@ncbi.nlm.nih.gov.

(**07/27/2017**)

## Allow up to Two Days for New or Changed eRA Commons Account Information to Become Effective

**I recently created an eRA Commons account and successfully logged in to my eRA Commons account. But when trying to login to the dbGaP system, I got an error message. Please help!**

The new account information or any changes to an existing eRA Commons account, such as user name, password, and email address changes, may take one to two days to be propagated from the eRA to the dbGaP system. After one or two days, please attempt to log in to the dbGaP Authorized Access System again.
(**07/27/2017**)

\

# Authorized Access System Login

Created: October 21, 2008; Updated: December 11, 2013.

## Login Page Redirect

**When trying to access the dbGaP log in page, after clicking on the "Login" link, I am redirected to a NIH login page. Does it mean something is wrong?**

The "Login" link on the dbGaP Authorized Access System login page is supposed to redirect you to a NIH login page where you can use the login credentials of your eRA Commons account to access the dbGaP system.

(**07/03/2017**)

## Troubleshoot Login Problems

**I am not able to login to the dbGaP system, please help!**

The dbGaP Authorized Access System authenticates users using NIH eRA Commons account information. **Your dbGaP login only works if your eRA Commons account login works.** Most of dbGaP account login problems are due to issues related to eRA Commons account login. **If you have problems logging in to your eRA Commons account, please seek assistance from the** eRA Commons helpdesk **before trying to login to the dbGaP system again.** If your eRA login succeeds but dbGaP login fails, please contact the dbgap-help at dbgap-help@ncbi.nlm.nih.gov.

(**07/27/2017**)

## Reset Login Passwords

**I lost my dbGaP user name and password. Can you please send me this information?**

The dbGaP Authorized Access System authenticates users using NIH eRA Commons account information. If you have a problem with logging into the dbGaP system, it could be simply because the login credentials (username or password) that you are using are incorrect. The login credentials for dbGaP login is the same as those of your eRA Commons account. You can reset your password online at the eRA commons "Reset Password" page.

After resetting the password and making sure it works for the eRA Commons account, please allow one or two days for the change to be propagated from eRA to the dbGaP system before trying to log in to the dbGaP system again.

(**07/27/2017**)

## New Password Works at eRA but not at dbGaP

**I recently created an eRA account and can use my username and password to log in to the eRA Commons system, but get an error message when I use them to attempt to log in to the dbGaP system.**

Please see here for the answer (**07/03/2017**)

# Data and Other Information Related to Controlled Access

Created: October 21, 2008; Updated: December 11, 2013.

## Authorized Access Related Information

**Where to find the information related to Authorized Access of a study?**

The information related to the Authorized Access can be found from the "Authorized Access" section on study page of the dbGaP public website. From the study page (such as this one), there is a link named "Authorized Access" right under the "Variables" tab on top of the page. The link is a shortcut to the section. Please see the screenshot of the section below.

The following information can be found from the section:

1. Email address of Data Access Committee (DAC) of the study
2. Data Use Certificate (DUC) of the study.
3. The name of the consent group.
4. Data use restriction of the consent group.
5. IRB requirement of the consent group.
6. The link to study-report and data manifest for the data available through the Authorized Access System.

Here is a real example of the section. You need mouse over the help ('?') icon to see the consent language of each consent group. (**09/27/2012**)

## IRB Requirement

**Where can I find the information about IRB requirements of dbGaP datasets?**

Please see here for the answer. (**07/03/2012**)

## Data Use Restrictions

**Where can I find data use limitation information of a study?**

The dbGaP data are organized into consent groups which consist of all the data from study participants who have agreed to the same data use as specified in the informed consent for the study. Data access is only approved in unit of consent group, and, therefore data requestors should understand the Data Use Limitations (DULs) of a consent group prior to applying for dbGaP data access.

DULs of a dbGaP study is listed in the "Authorized Access" section on the study page of dbGaP public website. From the study page (such as this one), there is a link named "Authorized Access" right under the "Variables" tab on top of the page.The link is a shortcut to the section. The DULs of each consent group is listed in the section. A screenshot of the section can be found here. (**07/27/2017**)

## Data Use Certification (DUC)

**Where can I find the Data Use Certification (DUC) for a particular study?**

The terms and conditions of using dbGaP data vary by study. A link to the Data Use Certification (DUC) of a particular study can be found in the "Authorized Access" section on the study page of dbGaP. From the study page (such as this one), there is a link named "Authorized Access" right under the "Variables" tab on top of

the page. The link is a shortcut to the section. A link to the DUC is included in the section. A screenshot of the section can be found from here. (**09/24/2017**)

## Data Manifest

**Can I get a list of data files available for download prior to submitting a Data Access Request (DAR)?**

A detailed description of distributed files can be found from the Study Configuration Report and Data Manifest available on the dbGaP public FTP site such as here. The FTP folder can be found through the link named "List of components" in the "Authorized Access" section of the study page. A screenshot of the section can be found here.

Please note that the information of SRA (Short Read Archive) data files is currently not included in the study report and data manifest. The SRA data here is a general term for the Next-Gen sequence raw data, BAM files, or other high throughput data. You may have to directly check the SRA or Biosample websites for the information. (**08/30/2012**))

## Data Manifest of Sequencing (BAM, SRA, etc.) Data

**I don't see BAM and SRA data in the data manifest available from the dbGaP FTP site. Where can I find the information?**

The sequencing data (including Next-Gen sequence raw data, BAM files, or other high throughput data.) distributed through the dbGaP are subject to frequent real time updates, the information of which, therefore, is not included in the study report and manifest available on the dbGaP FTP site.

The information of all data files available for download through the dbGaP Authorized Access System, including that of sequencing data files, however, can be found from a manifest file inside of the Authorized Access System. It can only be accessed through an approved Data Access Request (DAR). The following is how to locate it:

Login to the dbGaP account, go to "My Request" tab, and click on the "Request data" link in the "Actions" column of respective table row. There is a link named "Dataset Manifest" right above the "Create n ew data request for download" section. Clicking on the link allows saving or opening of a spreadsheet file. All the phenotype, genotype, and sequencing files are listed along with their sample ID information in the file.

(**07/14/2017**)

# Applying for Controlled Access Data

Created: October 21, 2008; Updated: December 11, 2013.

## Overview

**How does the dbGaP data access request (DAR) process work? Could you give an overview?**

The following is a general overview of how the dbGaP data access request (DAR) process works:

1. Applicants log in to the dbGaP Authorized Access System as a Principal Investigator (PI) using the username and password of the applicant's eRA Commons account. The applicant creates a project and follows multiple steps to complete and submit the online application. Note: requests for several consent groups from different dbGaP studies can be included in the same project. The application then is forwarded to an Institutional Signing Official (SO) for approval.
2. The SO approves the requests and they move into the queue of a NIH Data Access Committee (DAC) to review. The review is to confirm that the proposed research use(s) is consistent with the data use limitations (DULs) of the requested dataset(s). These restrictions are established by the language in each study's informed consent materials, and therefore vary across studies in dbGaP.
3. Once the DAC review is complete, the PI and SO is notified by email of approval or disapproval of the data access request (DAR).
4. Once the approval email notification is received, the PI can download the data by logging in to the same dbGaP account used for making the DAR.

A more detailed explanation of the DAR process can be found on the "How does one apply" section in the dbGaP Authorized Access system. It can also be found on this archive here.

Additional information about NIH policies and procedures for Genomic Data Sharing (GDS) Policy can be found on the NIH Office of Science Policy (OSP) website or you can send an email directly to gds@mail.nih.gov.

(**07/2272017**)

## Information of Data Available for Download

**Where can I find the information about the data files available for download prior to making a data access request (DAR)?**

Phenotype variables and datasets can be browsed through the public dbGaP web pages. Here is how to find the study report and data manifest that include the information of the phenotype and genotype data files.

Please note that the information of sequencing data (including Next-Gen sequence raw data, BAM files, or other high throughput data.) is not included in the study report and data manifest available from the public FTP site. That information, however, is available inside of the dbGaP Authorized Access System. Please see here for more information.

(**08/30/2012**)

## SRA Data Availability

**I've been approved to access all the data of the study, but I find that only the phenotype data are available for download. I know that the study is supposed to have SRA (Short Read Archive) data. Could you explain why the SRA data are not available?**

The dbGaP phenotype and genomic data are submitted and processed separately from the SRA data. The SRA data may include the Next-Gen sequence raw data, BAM files, or other high throughput data. It is quite common that the data submitter submits phenotype and genomic data of a study while publishing the paper, but holds on to the submission of SRA data due to either technical or other reasons.

The dbGaP often goes ahead and releases the study as soon as the phenotype and genomic data are ready independent of the SRA data. Once the study is released, the study page is published on the dbGaP website and the individual level phenotype and genomic data are made available through the dbGaP Authorized Access System. The SRA data release may fall behind either because the Principal Investigator (PI) may not have submitted the SRA data or the National Center for Biotechnology Information (NCBI) has not completed the data processing.. In cases where the SRA data lags as a result of the PI, the delay of SRA data release could be significant and sometime indefinite.

The SRA data submission and processing are handled by NCBI's SRA group. If the dbGaP study has been released and the SRA data have not yet been made, it is suggested to directly contact the sra@ncbi.nlm.nih for any questions regarding the status of SRA data release. They are in a better position to give you an update.

As the last resort, you may want to directly contact the PI of the study for more information. For most of studies, the PI's contact information is posted on the "Study Attribution" section at the bottom of the study page on the dbGaP website.

(**08/29/2017**)

## A Brief Step-by-Step for Filling Out the Data Access Request

**Would you give me a step-by-step instruction of how to apply for dbGaP Controlled-Access data?**

There are several videos available on YouTube that demonstrate the dbGaPdata access equest (DAR) process. The links to them are listed here.

The following is a brief step-by step guide to the dbGaP Authorized Access application process:

Visit the dbGaP Authorized Access System. When you get there, you will find that you will need an NIH eRA Commons account username and password to log into the system ().

To get an eRA account and password, go to eRA Commons, and follow the directions for online account registration. Please also see here for more about an eRA Commons account; For NIH researchers, in order to obtain access to genomic data in dbGaP that is available through controlled-access, eligible NIH Institute and Center (IC) intramural scientists and IC extramural program scientific staff must first obtain the approval of their IC. Please see here on additional information.

When you login for the first time, you will be asked to provide and save your contact information. This contact information will be used by the system to complete the cover page forms of your DAR. Once this is complete, you will not have to fill it out again in order to access the system in the future. Should your contact information change in the future, you can click on the "My Profile" tab located near the top of the page and update the information accordingly.

Click on the "My Projects" tab. Once you are on the "My Projects" page, click the "Create a new project" button to start a ata ccess equest. **Please be aware that the name of a new project cannot be the same as that of any existing projects, including closed projects.** After entering the information, move on to the next step using the "Begin New Research Project" button.

.

**Choosing Datasets**
All datasets available for download are listed under "Choose Datasets" sub-tab.

(Note: For a newly created project, this is the first sub-tab. After the first time, you will see this sub-tab before the "Confirm Dataset" sub-tab)

With hundreds of datasets listed on this page, it is difficult to find anything specific by browsing through the page. It is suggested to search the page using the web browser's search function (Control-F) to find related datasets. For example you will find Framingham datasets by searching for "phs000007" (no quotes) or by "Framingham".

Please note that study participants are partitioned according to their informed consent restrictions on use. Select only those participant sets that have consent restrictions consistent with your proposed research. **Please also be aware that if there is more than one consent group in a study, the participants of different consent groups do not overlap,** which means that you would have to get the data of all consent groups to cover all participants of the study.

Before including the datasets in the project you are creating, please read the "Data Use Limitations" of each dataset to make sure it matches the Research Use Statement that is required in a later step.

Select all the datasets you have decided to include using the checkbox and move to the next step by clicking on "Add Selected and Continue" button at the bottom of the page.

**Research Project**

On this page you will enter the following information

a. A title for the research activity
b. Type of research
c. A research use statement (limited to 2000 characters)
d. A non-technical summary (limited to 1100 characters)
e. The name of your Institutional Signing Official (SO)
f. The organization information of data requester or the Principal Investigator (PI).
g. Create decryption password for the project. The password is used to create dbGaP repository key that is required for configuring the NCBI decryption tool and SRA-toolkit. The decryption password and dbGaP repository key are shared among all download packages created under the project.

   Note: after being created for the first time, the decryption password can be reset from the page under the "Data Security" sub-tab. **Reset password will only affect the download packages created after the reset,** which means that the download package created before the reset will still use the old dbGaP repository key for decryption.

Move to the next step by clicking on "Save and Continue" button.

**Collaborators**

On this page, you will enter the names and contact information of your collaborator(s). The collaborators within your institution (if any) should be provided. The degree of detail for your list of collaborators is decided by your DAC (Data Access Committee) reviewers, but generally speaking, a "collaborator" is meant to include staff with an official appointment at your institution, and not supervised students and technical staff. Use the "add another collaborator" button if you have more than one local collaborator. The data downloaded from the dbGaP can be shared between listed internal collaborators through a secured computer system.

If investigators plan to collaborate with investigators outside their own institution, the investigators at each external site must submit an independent DAR using the same project title and Research Use Statement, and if using the cloud, Cloud Computing Use Statement. **The "Research Use Statements" section of each**

**application should also mention the respective external collaborators and the fact that this is a joint research project** so that the Data Access Committee (DAC) can review the requests together. These are important for an expedited review of separately submitted requests from all external collaborators.

**Information Technology (IT) Directors**
On this page you will enter the name(s) and contact information for your IT Director(s). Generally, a senior IT official with the necessary expertise and authority to affirm the IT capacities at an academic institution, company, or other research entity. The IT Director is expected to have the authority and capacity to ensure that the NIH Security Best Practices for Controlled-Access Data Subject to the NIH GDS Policy and the institution's IT security requirements and policies are followed by the Approved Users.

**Choose Datasets**
This is the same sub-tab mentioned early in this instruction. When the dataset is selected for the first time, this is the first sub-tab. After that it is placed to the middle before the "Confirm Datasets" sub-tab. Please see the early "**Choose Datasets**" section in this instruction for details.

**Confirm Datasets**
On this page, the datasets you selected in the "Choose Dataset" step are listed. This is your first chance to review the "Data Use Limitations" (DUL) of each selected datasets. Please make sure it matches the data uses described in the Research Use Statement.

For each listed dataset, if you want to remove it from the list, you can use the dropbox on the right side of the dataset name to select "Remove", and then use the "Remove Selected" button to remove it from the list. Please note that a datasets is no longer removable if it has ever been approved.

**Review the Data Use Certification**
On this page, you are given a chance to review the Data Use Certificate (DUC ) and the Dataset Manifest (look for PDF icons in the right most column) for each of the datasets you have requested.

- At this point you can use the "Back" button to go back to the previous step if you want to remove datasets or go back to the "Choose Datasets" step if you want to add datasets.
- Review the DUCs and the Dataset Manifests to make sure everything is correct.
- If any of the datasets included requires IRB approval, go to the "Program-specific required attachments" section at the bottom of the page

(**Please note that the "Upload" or "Delete" button may not be shown if no dataset is selected at the "Choose Dataset" Step**).

The IRB requirement is imposed by the data submitter. For the datasets marked as IRB required, the IRB approval document has to be provided in order for the system to go through. There are no other alternatives.

**Review the Data Use Limitation**
On this page, you are given a final opportunity to review the Data Use Limitation (DUL) of each selected dataset and make sure it matches the data uses described in the Research Use Statement. If the DUL of any of selected dataset doesn't match intended data use, you would need to go back to previous "Confirm Datasets" step to remove it. If you agree the DUL are consistent with your intended use as described in your Research Use Statement, make sure all the checkboxes are checked and move to the next step by "I agree the Data Use Limitation(s)" button.

**Review Applications (& Electronic Signature)**
On this page, you can find links to your completed application(s). You may want to review your application(s) as a whole before the final submission. If you need to revise any portion of your applications, use one of the sub-tabs to return to a previous step and make changes.

Once you have thoroughly reviewed the application,, you will need to click the checkboxes labeled "I agree

…". **Checking these two checkboxes constitutes an electronic signature,** which affirms the accuracy of the application. After clicking the two check boxes, click the "Submit Application to Signing Official" button to route the request to your Institutional Signing Official (SO) for their signature.

After the application being submitted, under "My Requests" tab, you can see that selected data requests of the project are under "SO review" status. In the meanwhile, an automatically generated email notification will be sent to the designated SO for reviewing the application. Once being approved by the SO, each DAR under the project will be in queue for review by the respective Data Access Committee (DAC). The status of the request is changed to "DAC review".

After the application is submitted and before being pick up from the SO queue, you can use the "Recall Application" button to withdraw the submission. After the recall, you can revise and submit it again.

Please see here for how to check the status of submitted DARs.

You will be notified once DAC review is done. You will be allowed to download requested data once the DAR gets approved by the DAC.

(**7/11/2017**)

## Check Status of Submitted Data Requests

**How do I check the status and review my data access requests (DARs)?**

You can track the progress of your submitted data access requests (DARs) through the dbGaP Authorized Access System. After logging in, click on "My Requests" tab located near the top of the page, which takes you to the "My Requests" page.

The progress of your DAR is indicated in the "Status" column of the request table. The following are a few commonly seen statuses:

a. **SO review**: After the dbGaP project is submitted, the DAR(s) under the project will be under "SO review" status. In the meanwhile, an automatically generated email will be sent to notify the Signing Official (SO) to review the data requests. You may want to directly contact the SO to expedite this process.
b. **DAC review**: After being approved by the SO, the DAR is put in the queue for an NIH Data Access Committee (DAC) to review. The time required for DAC review varies depending on many factors, however, average DAC review is two weeks. Please directly contact the respective DAC for an update if necessary.
c. **Approved**: Once the request is approved by the DAC, you will receive an email notification about the status of the request. The status of the request will be shown as "Approved". Once approved, you can immediately start to assemble download packages by going to the DAR page through "My Requests" tab and download the data through "Downloads" tab.
d. **Approved Expired**: Through submission of the DAR, the Principal Investigator (PI) agrees to submit either a project renewal or close-out request prior to the expiration date of the 1-year data access period. Failure to submit a renewal or to complete the close-out process, including confirmation of data destruction by the Institutional Signing Official, may result in termination of all current data access and/or suspension of the PI and all associated personnel and collaborators from submitting new DARs for a period of time..
e. **Rev. Requested**: If the DAR is under "Rev Request" status, please go to the data request page through "My Request" tab and read DAC comment. You may need to revise the project accordingly and resubmit it through your dbGaP account. Please see here for more about how to revise and resubmit an existing dbGaP project.

f. **Rejected**: If your DAR is rejected by the DAC, it often means that stated data use does not match the data use limitation of selected datasets. Please go to the DAR page through "My Request" tab and read the DAC comment for more details. You may need to reselect datasets, change your Research Use Statement, or make other necessary changes, then resubmit the project through your dbGaP account. Please see here for more about how to revise and resubmit an existing dbGaP project.
g. **To review**: After a project is submitted for "SO Review," if by some reasons the application is returned back to the PI by the SO for revision, or the PI has chosen to withdraw the application using "recall from SO" link, or "Recall Application" button, all the DARs under the project will be under "to review" status. This gives the PI an opportunity to revise and resubmit the project for SO review again.

(**7/13/2017**)

## Time Estimate for Data Access Committee (DAC) Review

**I've applied for two studies. Can you give me an estimate for how long it will take for a decision to be reached after the application is submitted?**

The length of time for processing an application request depends on:

1. How fast the Signing Official (SO) signs the application.
   Through an automatically generated email, the SO of the dbGaP project should be notified to review the data access requests (DARs) when a project is submitted. You may want to directly contact the SO to expedite this process.
2. How fast the Data Access committee(s) DAC approves the application.
   The time required for DAC review varies depending on many factors, however, average DAC review is two weeks

**Since dbGaP is not directly involved in the review and approval process of DARs, you may have to directly contact respective DACs to check on the progress of your application.** You will find that your application includes the email address for the DAC of each study. The DAC email address can also be found from the "Authorized Access" section of respective study page on dbGaP public website (see here for the screenshot).

(**07/05/2017**)

## Signing Official (SO) and IT Director

**How can I find SO and IT Director for my dbGaP project?**

A list of the Signing Official (SO) at a Principal Investigator (PI's) institution registered with the NIH eRA Common is available under the "Project details" page. To find the page, go to "My Projects" tab, click on the project name, and click on "Project details" sub-tab. The SO registration with the eRA is often handled by institution's eRA administrator. Please see here for more about institutional Signing Official (SO).

The IT Director is generally, a senior IT official with the necessary expertise and authority to affirm the IT capacities at an academic institution, company, or other research entity. The IT Director is expected to have the authority and capacity to ensure that the NIH Security Best Practices for Controlled-Access Data Subject to the NIH GDS Policy and the institution's IT security requirements and policies are followed by the Approved Users..
(**02/06/2013**)

## Select Datasets by Consent Groups

**How to select dbGaP datasets? Where can I get the consent information of a study prior to submitting a data access request?**

The dbGaP data are organized into consent groups which consist of all of the data from study participants who have agreed to the same data use limitations as specified in the informed consent for the study. When applying for dbGaP data, the data available are grouped in unit of consent groups.

Data access is only approved in unit of consent group, the data requesters therefore should understand the Data Use Limitations of a consent group prior to applying for dbGaP data access. Data Use Restriction of a dbGaP study is listed in the "Authorized Access" on study page. Here is how to find it.

Please note that, **for a given study, participants in different consent groups do not overlap**, which means that you would have to get the data of all consent groups to cover all participants of the study. (**7/27/2017**)

## No Need to Complete the Request in a Single Session

**Is it possible to begin a data access request; complete some sections, save the work and come back later to complete further sections?**

Yes, you can complete some sections of the application and return to finish in one or more additional sessions, as the data access request application system will save your application at each step when you click the green "Next Step" button at the bottom. Note: the text of the green button may be slightly different depending on the context of the button. (**08/11/2017**)

## How to Write the Research Use Statement

**What kind of research use statement is necessary for a data access request? Does my project need to be closely related to the dbGaP study of interest to qualify for access?**

NIH Data Access Committees (DACs) review research use statements to make sure the proposed research is consistent with the data use limitations on the requested datasets in the data access request(s) (DAR).the RUS should include the following components:

- Objectives of the proposed research;
- Study design;
- Analysis plan, including the phenotypic characteristics that will be evaluated in association with genetic variants;
- Explanation of how the proposed research is consistent with the data use limitations for the requested dataset(s); and a brief description of any planned collaboration with researchers at other institutions, including the name of the collaborator(s) and their institution(s). Additional guidance to how to write a research use statement can be found on the Office of Science Policy website..

Some studies have additional instructions that are required in their applications for controlled-access. If you have study-specific questions, you may wish to contact the Data Access Committee (DAC) for that study. The email address for each study's DAC is provided in the "Authorized Access" section of the study page (see screenshot here ).

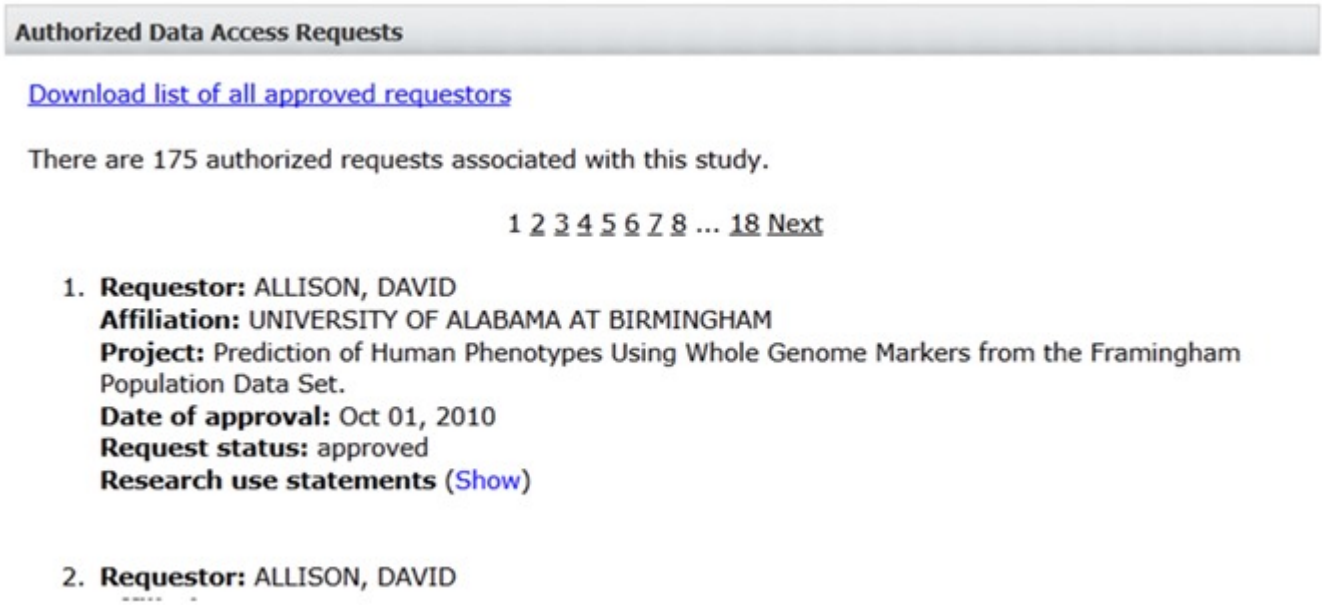
(**10/19/2017**)

# Examples of Research Use Statements

**I am working on the online application forms. Are there examples of research statement of approved data requests that I can take a look?**

All research statements of approved data access requests (DARs) are posted on the respective study page of the dbGaP public website. There is a link named "Authorized Requests" on top of the study page (see screenshot below).

The link leads to "Authorized Data Access Requests" section that contains research use statements of approved datasets. You may need to use "Show/Hide" next to "Research Use Statement" on each requester's section to view text. The "Public Research Use Statement" and "Technical Research User Statement" are both included.




**07/03/2012**)

# IRB Requirement

**What is IRB approval? How do I know if a specific dataset needs IRB approval to access?**

An Institutional Review Board (IRB), privacy board, and/or equivalent body is a committee that reviews and monitors research involving human subjects. The group serves an important role in the protection of the rights and welfare of human subjects research. Some datasets require IRB, privacy board, and/or equivalent body approval for use, as noted on the dbGaP study page displayed in the "Authorized Access" section.. Please see here for more about how to locate it.. Other types of documentation may be required as described on the study page and/or Data Use Certification for the study. Evidence of IRB approval and other documentation can be uploaded as a PDF during the application process.

(**08/01/2013**)

## For Institutional Review Board (IRB)-Related Questions Contact the Data Access Committee (DAC)

**The requested dbGaP dataset requires IRB approval, however, we do not belong to a university, and thus do not have a local IRB. Will approval from a non-local IRB be acceptable?**

Any questions about IRB approval should be directed to the Data Access Committee (DAC) for the study in question. The email address of the DAC for a specific study is in the "Authorized Access" section of the Study page. Please see the screenshot at here. (**06/15/2011**)

## Add or Update Institutional Review Board (IRB) Approval Documents

**How to add or update IRB approval documents for my data requests?**

If any of the datasets included in a dbGaP project have an IRB requirement, the respective IRB approval documents in a PDF format will have to be uploaded in order to complete the application process. This needs to be done by the Primary PI of the data access request (DAR) through the dbGaP Authorized Access System. The following is how to add or update IRB documents:

After logging in to the dbGaP system, please do the following:

1. Click on the "My Projects" tab and click on the project name.
2. For new projects or for existing projects that wish to add datasets, make sure datasets are selected under "Choose Datasets".
3. Confirm selected datasets by hitting "Next" button until you reach, "Review DUC".
4. On the "Review DUC" page, if any of the datasets included in the project have an IRB requirement, you will see the "Program-specific Required Attachments" section at bottom of the page. Use the "Upload" button to upload IRB approval documents for the datasets shown in the table.
5. To update previously uploaded IRB approval documents for an existing project, go to the "Review DUC" page, in the "Program-specific Required Attachments" section at bottom of the page. Use the "Delete" button to remove existing IRB approval documents and use the "Upload" button to upload new ones.

If you don't see the "Upload" button, please be aware that the **"Upload" button is visible only if the dataset that requires an IRB approval document is selected.** You may want to go back to the "Choose datasets" page to select the dataset first.

(**10/19/2011**)

# Revise, Amend, and Update Existing Application

Created: October 21, 2008; Updated: December 11, 2013.

## Revise Data Access Request

**How to make changes to my existing dbGaP project?**

A dbGaP data request is organized by project. A project could contain multiple data access requests(DARs) from one or more dbGaP studies. Thus, any change made to one or more DARs under the project necessitates that the entire project has to be revised.

Most steps of project revision are simply repeating those that occurred while creating a new project creation. Revised and resubmitted projects will go through SO review and subsequently be in the DAC queue for review again.

**Please note that approved DARs will stay as approved during the re-review process, unless the revision is ultimately rejected by the DAC(s).**

The following are some common situations where revision and re-submission of a completed project is needed:

a. Change user profile
   Note: the profile items include phone number, position/title, and user address. Principal Investigator's (**PI's) email can only be changed directly through the eRA**.
b. Change project information or data use statement.
c. Change collaborator or IT Director.
d. Change Signing Official (SO).
e. DAR is rejected or sent back for revision by the Data Access Committee (DAC).
f. Add or delete datasets from project.
g. Add or update IRB documents.

The following is a detailed instruction.

Log on to the dbGaP Authorized Access System as a Principal Investigator (PI).

Go to the "My Projects" tab if you are not already there.

Note: **If any DAR under the project is currently under "SO Review", the project has to be recalled from the SO first before the revision can be made.** The status of the DAR can be found under the "My Requests" tab and here is how to recall a project from the SO.

Go to the project page by clicking on the "Revise Project" link from the "Actions" column of the project that needs to be revised. Another way of getting to the project page is to simply click on the project name under "My Projects" tab.

If the "revise project" link is not in the "Actions" column, it may suggest that the year-end renewal date of the project is approaching or passed. In this case, you should instead click on the link named "Renew Project" to start the project renewal process. Please see here for more about it.

**Research Project**
Make necessary changes to the information presented on the page. This may include data use change, new information related to revision and update etc. The change of the Signing Official (SO) can also be made on this page.

**Collaborator**
Change internal and external collaborators information if necessary. For external collaborators, the information of related dbGaP requests should be included. Please see here for more about it.

**IT Director**
Change IT Director information, if necessary.

**Choose Datasets**
You can add datasets not currently included in the project at this step. You may have to reselect the rejected datasets in the project if you want to include them again.

To navigate through hundreds of datasets listed on the page, it is suggested to search using the web browser's search function (Control-F) to find related datasets. For example, you will find Framingham datasets by searching for "phs000007" (no quotes) or by "Framingham".

If any datasets are selected, use the "Add Selected and Continue" button to move to next step.

You would need to go through the rest of the steps listed below) to re-submit the project.

**Confirm Datasets**

**Review DUC**

**Review DUL**

**Review Applications (& Electronic Signature)**

The detailed description of steps can be found in the creating a new project section..

The submission will go to the SO for re-review and an automatically generated email notification will be sent to the SO. If approved by the SO, the DAR will be in the DAC queue for reviewing again. You will be notified once the requests are approved by the respective DAC.

(**12/11/2013**)

## Recall is required before Making Changes to Project under SO Review

**I am not able to make changes to my project that has data access requests that are under "SO Review." What do I need to do?**

If a project contains any data access requests (DARs) under "SO Review", in order to make changes to the project, the Principal Investigator (PI) will need to either contact the Signing Official (SO) with a request to return the application for revision or use the link named "Recall from SO" to withdraw the application from the SO queue. The following is how to make the recall.

1. Login to the dbGaP Authorized Access System.
2. Go to the "My Projects" page (this is default page you are presented upon login).
3. Click on the link "Recall from SO" in the "Action" columns of the project. You should be redirected to the "Review Applications" sub-tab of the project page.
4. Click on the "Recall Application" button. A text box for a comment should be displayed.
5. Enter the comment and click the "Recall Application" button again.
6. At this point, you are ready to make any necessary revisions or reselect datasets through respective sub-tabs, and finally resubmit the project to the SO for review again. Please see here for a more detailed instruction of how to revise your project.

(**08/06/2013**)

## Changing Signing Official

**How do I change the Signing Official (SO) designated to my dbGaP project?**

If any of the data access requests in the project is under "SO Review", you need to recall the project from the Signing Official (SO). Please see here for more details.

After being recalled from the SO or if the Data Access Request has been completed and approved by the current SO, you can change the SO of your project by the following steps:

1. Log into the dbGaP Authorized Access System. Click on the "My Projects" tab if you are not on the "My Research Projects" page.
2. For the project to which you need to change the SO, click on the project name, or from the "Actions" column of the project, click on the "Revise Project" link. This will take you to the project page.
3. Click on the "Research Project" sub-tab and choose from the list of SOs of your organization registered with eRA Commons.
4. Go through of the remaining steps and submit the change.

The new SO will be informed to review the submitted request.

It often happens that the new SO intended to be designated for the project is not found from the SO list registered by the primary PI's organization in the steps described above. In this case, the PI's organization will have to first register the new SO with eRA Commons before the PI can make the SO change. Please beware that any changes made to the eRA system may take one or two days to become effective in the dbGaP system. Please contact the eRA administrator of your organization or visit the eRA Commons website for more about how to add an institutional SO. The eRA Commons How To and FAQ sites are often found to be useful.

(**08/02/2013**)

## Changing the Principal Investigator (PI) for Project

**How do I transfer my project to another PI? (PI to PI transfer)**

Process for Transferring Principal Investigators Within the Same Institution:

Step 1: Principal Investigator (PI) selects action transfer to another PI.

Step 2: PI fills out online form to transfer application.

Step 3: PI submits transfer application to Institutional Signing Official (SO).

Step 4: SO certifies transfer application.

Step 5: Email alert is sent to an NIH Data Access Committee(s) informing them of transfer

Step 6: dbGaP Project application shows up in new PI's account and is removed from the previous PI's account.

(**05/03/2019**)

## Transferring Data Request to a Different Institution

**I am about to leave my current job and start a new one in a different institution. Can I transfer my approved data requests to the new institution?**

The Signing Official (SO) and IT Director of the Principal Investigator's (PI's) institution are ultimately responsible for enforcing proper data use and data security. The dbGaP data applications therefore are not transferable to different institution. Prior to leaving a current institution, the PI needs to close out current dbGaP project and file a new application with the new institution (see here for more about project closeout). It may be helpful to directly contact the Data Access Committee (DAC) that approved the request to see if there is an expedited review given the situation (**09/23/2011**)

# Adding New Lab Members from the Same Institution

**I have new people in my lab that arrived after I received approval for the data sets I requested. How do I amend my data access request to include them?**

The degree of detail for your list of collaborators is decided by your DAC reviewers, but generally speaking, a "collaborator" is meant to include staff with an official appointment at your institution, and not supervised trainees such as graduate students or postdoctoral fellows.

The step-by-step process for adding collaborators to a data access request (DAR) is as follows:

1. Navigate to the "My Projects" page in the dbGaP authorized access system.
2. Click on the "Revise Existing Request form(s)" link for the project to which you need to add personnel (the link is located below the title of the project).
3. Select Step 2a (Collaborators) from the grey tabs at the top of the page.
4. Enter the name(s) and contact information for your new lab member(s) and press "Save" or "Next" button
5. Use the "Next" button at the bottom of each succeeding page until you get to Step 6.
6. At step 6, re-submit your application

Please see here for more about collaborators (**06/21/2011**)

# Requested Wrong Data Sets

**Can I remove previously selected dataset from my project?**

You can delete unwanted datasets included in your project before the data access request (DAR) is approved by the Data Access Committee (DAC). The dataset is no longer removable from the project once it is approved for access. Please refer to the project revision section at here for more details. (**08/07/2013**)

# Change User Profile

**How do I change my dbGaP account user profile?**

The user profile of the Principal Investigator's (PI's) dbGaP account can be changed through the "My Profile" tab. Please note that the grey field values are automatically filled in by pulling in the information from the PI's eRA Commons account. These fields include first and last names and email address. The eRA Commons account information can only be changed directly through the eRA system. Please visit eRA commons website for information. The eRA Commons How To and FAQ sites are often found to be useful.

Please be aware that any changes to an eRA Commons account may take one to two days to be propagated from eRA to dbGaP system.

(**08/02/2012**)

# Contact Information

Created: October 21, 2008; Updated: December 11, 2013.

## Contacting dbGaP

**I have several questions regarding an application to dbGaP. What e-mail address should I send my inquiry to?**

dbgap-help@ncbi.nlm.nih.gov is the correct email address for questions about the application process. General questions about the NIH Genomic Data Sharing (GDS) policy should be addressed to gds@mail.nih.gov.

(**08/22/14**)

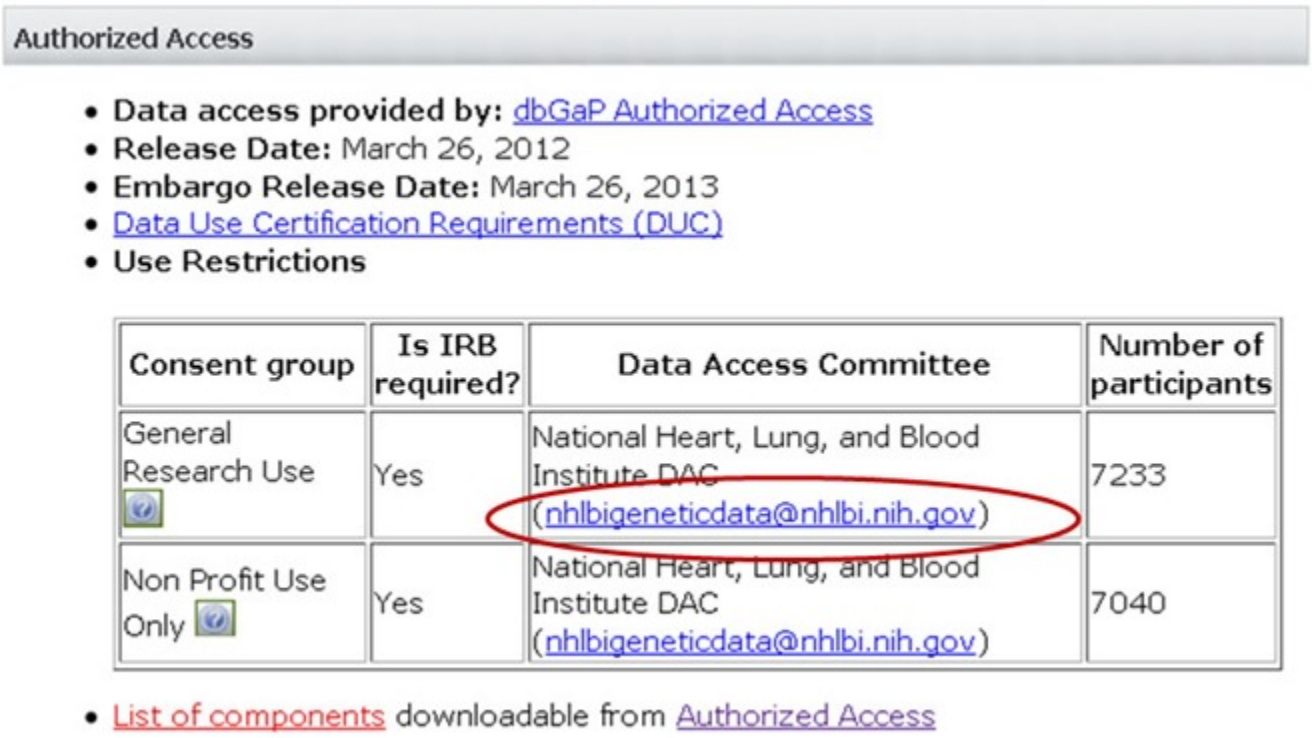
## Contacting Data Access Committee (DAC)

**Where can I find the DAC (Data Access Committee) email address of a study?**

The DAC email address can be found either in the dbGaP Authorized Access System or on the dbGaP website.

For the former, once logging in to the dbGaP account, go to the "My Requests" tab, the DAC name (such as NHGRI, GAIN, TCGA …) is listed in the "Data set" column of the data request table. Click on the name, the email editor with the DAC email address will be opened.

On the dbGaP website, the email address of the Data Access Committee (DAC) for a specific study is located in the "Authorized Access" section of the Study page (**circled in red below**):



Authorized Access

- **Data access provided by:** dbGaP Authorized Access
- **Release Date:** March 26, 2012
- **Embargo Release Date:** March 26, 2013
- Data Use Certification Requirements (DUC)
- **Use Restrictions**

| Consent group | Is IRB required? | Data Access Committee | Number of participants |
|---|---|---|---|
| General Research Use | Yes | National Heart, Lung, and Blood Institute DAC (nhlbigeneticdata@nhlbi.nih.gov) | 7233 |
| Non Profit Use Only | Yes | National Heart, Lung, and Blood Institute DAC (nhlbigeneticdata@nhlbi.nih.gov) | 7040 |

- List of components downloadable from Authorized Access

(**08/29/2012**)

## Contact for Short Read Archive (SRA ) Related Questions

**I have some questions related to the Short Read Archive (SRA ) data distributed through the dbGaP. What e-mail address should I send my inquiry to?**

If your question is about SRA data availability or release status, please see here for more information. The SRA data may include the Next-Gen sequence raw data, BAM files, or other high throughput data. The SRA data distributed through dbGaP is handled and processed by the National Center for Biotechnology Information's (NCBI's) SRA group. The SRA group, thus, is in a better position to answer most of the SRA-related questions, especially those about SRA data submission, availability, and quality. The SRA help email address is sra@ncbi.nlm.nih.gov (**06/14/2011**)

## Contact for SRA Toolkit Related Questions

**I tried to install sra_sdk-2.1.10 but got some errors. Where should I look for help?**

The SRA toolkit can be downloaded from here. The best contact for any SRA toolkit related issues is sra-tools@ncbi.nlm.nih.gov, a mailing list documented in every README within the source downloaded by users.

(**07/10/2012**)

# Expiration Date, Renewal, Project Suspension, and Closeout

Created: October 21, 2008; Updated: December 11, 2013.

## Data Request and Project Expiration Date

**I received a reminder for expiration of my data access request a few days ago, why do I get another one? I wonder how the expiration date is defined.**

All data access requests (DARs) are contained in a project, which manages administrative information of the requests under the project. The approval of a DAR is typically granted for a one-year period. After initial request are made, more requests can be added to the project. Different DARs under a project thus can have different expiration dates. Consequently, **a Principal Investigator (PI) may receive multiple automatically generated email reminders for the expiration of different requests or for projects in a relatively short period of time**. The status of a DAR is shown as "Expired" after the expiration date. Here is how to check status of dbGaP data requests.

In addition to data request expiration date, there is also a project level year-end renewal date, through which the project information that governs all DARs in a project can be updated and reviewed once a year. When the year-end renewal date is approaching, the project needs to be renewed to stay as approved even there is no intention to make any changes. At the minimum, the "Data Use Statement" may be revised to indicate the future data use in the next approval period. **The PI's dbGaP account will be suspended if the year-end-renewal or project closeout request is not submitted 42 days after the project expiration date.**

Automatically generated email reminders will be sent to the PI 30 and 14 days prior to each of expiration dates mentioned above. **It is suggested to submit the renewal request as soon as the first email reminder is received,** which gives the DAC ample time to review it.

**The year-end-renewal of a project will reset the expiration date of the project and that of all requests under the project to the same new date.**

(**12/11/2013**)

## Renewal Procedure

**I've received an email from dbGaP that reminds me that the year-end renewal date of my dbGaP project is approaching. What is the procedure to submit a renewal?**

The renewal of data access requests (DARs) within a project requires the renewal of the entire project. The renewal process, however, will not affect currently approved requests. **The approved requests will stay as approved during the renewal process**.

(Note: **It is no longer required to send a separate annual report directly to the Data Access Committee**.).

**Procedure for renewal**

The following procedure is for the year-end renewal a dbGaP project.

Log on to the dbGaP Authorized Access System as a Principal Investigator (PI).

Click on the "My Projects" tab. From the project table, find the project that contains the data requests that you like to renew. Click on the link named "renew project" in the "Actions" column on right side. This leads you to the project renewal page, which starts from default sub-tab "Research Progress".

If you don't see the "Renewal Project" link in the "Actions" column, it means that it is too early for you to make the year-end-renewal. In this case, you may consider user the "revise project" link to revise the project. Please see here for more about it.

If there is a need to make changes to the information under sub-tabs before the sub-tab of the default page you are seeing, this is an opportunity to do so. Otherwise, you can start from the sub-tab "Research Progress", go through subsequent steps, and submit the renewal request at the end.

**Research Progress**
On this page you will enter the information about research progresses and intellectual properties resulting from analyzing requested data, and research plan for specified datasets. The datasets accessions and names of approved dataset pertaining to the research plan should be provided. Move to next page using "Save and Continue" button.

**Presentation**
On this page, you are required to provide information of presentations resulting from analyzing requested data. Please check the checkbox if you don't have any presentations. Move to next page using "Save and Continue" button.

**Publications and Manuscripts**
On this page, you are required to provide all publications resulting from analyzing requested data. Please check the checkbox if you don't have any publications. Move to next page using "Save and Continue" button.

**Data Security**
On this page, you are required to provide information related to data security. Please carefully follow the instructions on the page and provide the information as detailed information as possible. It is important for privacy protection of the individual level data. Move to next page using "Save and Continue" button.

The rest of steps are shown below, the instructions of which are identical to that of the respective steps of creating a new project.

**Choose Datasets**

**Confirm Datasets**

**Review DUC**

**Review DUL**

**Review Applications**

After project renewal is submitted, the status of the DAR will be changed to "SO Review", and the Signing Official (SO) of the requests under renewal will be notified by an automatically generated email. You may want to directly contact the SO to expedite the process.

After being approved by the SO, the DAR will be in the DAC queue for review and the status is changed to "DAC Review". Please check to make sure the status is changed. Here is how to check the status of dbGaP requests.

During the renewal process, **approved requests will stay as approved unless the renewal of entire project gets rejected by DAC**.

Please note that NCBI and dbGaP are not directly involved in DAR approval or the renewal process. For any questions concerning DAR approval or the renewal process, please **directly contact the DAC(s)** to get an update of the renewal status or to resolve any outstanding issues.

(**12/11/2013**)

## Annual Report

**Do I need to submit annual report directly the DAC when making year-end renewal of my project?**

**It is no longer required to send a separate annual report directly to the DAC**. The project renewal process now is handled entirely through the web-interface from the Principal Investigator's (PI's) dbGaP account. Please see here for more details.

(**08/08/2013**)

# Project Closeout

**The research related to my dbGaP project is no longer active. How do I precede to close out the project?**

A dbGaP project can contain one or more approved datasets. If the research of the project is no longer active, it is a good idea to close the project, so that the Principal Investigator (PI) will not have to make yearly renewal of the project. The data access request (DAR) expiration or year-end-renewal reminder will not be sent to PI.

**Procedure for project closeout**

Log on to the dbGaP Authorized Access System as a Principal Investigator (PI).
Click on "My Projects" tab. From the project table, find the project that contains the data request that you want to close. This leads to the "Project Details" sub-tab.

Read the project information carefully to make sure that it is the right project to closeout. Move the next page using "Begin Close Out Process".

**Research Progress**
On this page you will enter the information about research progresses and intellectual property resulting from analyzing requested data. Move to next page using "Save and Continue" button.

**Presentation**
On this page, you are required to provide information of presentations resulting from analyzing requested data. Please check the checkbox if you don't have any presentations. Move to next page using "Save and Continue" button.

**Publications and Manuscripts**
On this page, you are required to provide all publications resulting from analyzing requested data. Please check the checkbox if you don't have any publications. Move to next page using "Save and Continue" button.

**Data Security**
On this page, you are required to provide information related to data security. Please carefully follow the instructions on the page and fill in as detailed information as possible. It is important to provide complete information and faithfully report any incidents or issues you consider to be relevant to data security. Move to next page using "Save and Continue" button.

**Reasons for Project Closeout**
On this page, please state the reason of the closeout by checking appropriate checkbox. More than one reason can be selected. You can also describe the reason or provide additional comments in the text box. Move to next page using "Save and Continue" button.

**Review Closeout Application**
Before completing this page, upon project close-out, the PI and all approved users agree to destroy all copies, versions, and derivations of the dataset(s) retrieved from NIH-designated controlled-access databases, on both local servers and hardware, and if cloud computing was used, delete the data and cloud images from cloud computing provider storage, virtual and physical machines, databases, and random access archives, except as required by publication practices, institutional policies, or law to retain them.. The submitted closeout request is summarized in a PDF document provided as a link on the page. Please review the summary information and, if all is satisfactory, check the 'I Agree' checkbox. Move to the next page using "File Report and the "Close Project" button.

The request will be submitted to the Signing Official (SO) of the project for approval. The SO is required to confirm the data destruction and insure retained data is encrypted, properly stored, and deleted at the

appropriate time to comply with data security policies. If approved by the SO, the request will be sent to the Data Access Committee (DAC) for final approval. The project will be closed out after DAC approval is completed.

(**10/20/2017**)

# Account Suspension

Created: October 21, 2008; Updated: December 11, 2013.

## Why is My Account Suspended and How Do I Fix This Issue?

**I have just found out that my dbGaP account is suspended. Could you explain why and what I need to do to reinstate my dbGaP privileges?**

The dbGaP account could be suspended for many reasons, such as data use violation or late annual report. The most common reason is a result of overdue annual report. **The primary PI's dbGaP account will be suspended if the year-end-renewal or project closeout request is not submitted 42 days after the project expiration date.** Please see here for more information about the dbGaP project expiration date.

When an account is suspended, the PI can still login to the account and review all the information in the account. The functions of making new data request and data download are disabled. The reason for account suspension is displayed in a prominent position in the account. The decision of suspending a PI's account can also made by the related Data Access Committees (DACs).

If the suspension is because of an overdue year-end project renewal, the PI should renewal the project as soon as possible. Detailed instruction of project renewal can be found here.

If the suspension is because of an overdue year end project renewal and the PI wishes to close out the project, detailed instruction of the close out process can be found here.

(**10/20/2017**)

# Collaborators

Created: October 21, 2008; Updated: December 11, 2013.

## Collaborators within Primary PI's Institution

**I am working on the "Collaborators" section of the dbGaP data access request. What are the requirements for internal collaborators?**

You would need to provide the full legal names and contact information for all additional investigators from your institution who will have access to the dataset(s). (Exclude trainees, who are covered under the NIH policy). By submitting names on this form, requestors and signing officials guarantee that these individuals have read and agreed to the terms, conditions, and statements of the respective Data Use Certification(s).

Please note that collaborators from other institutions must submit a separate Data Access Request(s) for this project from their respective institution(s). All collaborators must be approved users before data can be shared. Coordinated requests by collaborating institutions should each use the same project title and should each complete this section in their respective applications.

(**03/10/2021**)

## Collaborators outside Primary PI's Institution

**I am an authorized user for controlled-access data. Can I add another investigator who is outside my organization to my dbGaP project?**

Yes, you can. You would need to provide the full legal names and contact information for external collaborators (i.e., those employed outside Coprimary PI's institution). External collaborators should be listed in the external collaborator(s) section of the project request applications. Data exchange between all collaborators must be consistent with the NIH Security Best Practices for Controlled-Access Data Subject to the Genomic Data Sharing (GDS) Policy and GDS Policy.

External collaborators must submit a project request with (1) the same project title and (2) a Research Use Statement and Cloud Use Statement, if applicable, that references the collaboration (for smaller collaborations, the name and institution of the collaborating PI(s) or for larger efforts, the consortium name)..

You can revise your application to include such a statement if it hasn't been included. Please see here for more about how to revise and resubmit a dbGaP project.

(**07/23/2013**)

## Collaborators who Contract with the Primary PI's Institution

**I'm university staff and will collaborate with a company whose contract is with my university. Should the company submit a separate data request form?**

The company that has contract with your institution is an external collaborator. They must submit a separate data access request since each data access request is specific to one institution: Please see here for more about external collaborators.

(**07/12/2012**)

## Add or Remove Collaborators

**How can I add or remove collaborators from my dbGaP project?**

To add or remove collaborators from a dbGaP project, the Principal Investigator (PI) needs to revise and resubmit the project through the dbGaP system. Please see here for more details of how to revise and resubmit a dbGaP project..

(**10/20/2017**)

# Signing Officials (SO)

Created: October 21, 2008; Updated: December 11, 2013.

## Lost Passwords

**I'm a Signing Official (SO) for my institution and forgot my password. Could you please assign me a new password?**

The login passwords are managed directly through the NIH eRA system. Please see here for how to reset password.

(**07/12/2012**)

## How to Sign Off on a Data Access Request?

**I'm the Signing Official (SO) for my institution and would like to sign-off on a Data Access Request but cannot find where I'm supposed to do this.**

1. Log into the dbGaP authorized access system using your eRA account login credentials.
2. As this will be your first time using the system, you will first be taken to a "Preferences" page where you will need to complete basic contact information (address, email address, etc.) needed by the dbGaP system. Once this is complete, you will not have to fill it out again to access the system in the future. Should your contact information change in the future, you can click on the "Preferences" link located in the cream-colored box located to the right of your "SO Projects" page, or click on the "My Profile" tab located near the top of the page.
3. After supplying the system with your contact information, you will be taken to the "SO Projects" page, which lists your queue of research applications needing review and approval.
4. To approve a request, simply click on the title of the request. You will then need to check the two boxes located in front of approval statements. Once this is done, click the "Approve and Submit to DAC" button. This will route the application to the appropriate NIH Data Access committee (DAC) for review.

(**06/17/2011**)

)

## Changing Signing Official

**The Signing Official (SO) currently assigned to the Primary PI's dbGaP project is incorrect. How can I change the SO?**

The Signing Official (SO) for a dbGaP project can only be made by the Principal Investigator (PI). Please see here for more about how to make the change. (**09/30/2011**)

# Downloading and Extracting dbGaP Data

Created: October 22, 2008; Updated: August 12, 2013.

This section of the dbGaP FAQ Archive contains information about how to request, download and extract individual-level data contained in dbGaP, as well as instructions for those with common request, download, and extraction problems.

**Please Note**: All the FAQs in this section assume that you have completed the dbGaP data access request process, and have been granted authorized access to individual level data. If you do not have authorized access to dbGaP individual-level data, accessing individual-level data using the instructions here will not be possible, as you will not have access to the appropriate password-protected dbGaP sites mentioned.

**To begin searching this section of the dbGaP FAQ Archive, you can either**:

- Enter your search word(s) text in the text box at the top of the page and click on the "Go" button,

  OR

- Click on any of the "Requesting, Downloading and Extracting dbGaP Data" sub-categories listed in the navigation box on the right side of the page to navigate to the sub-category of your choice.

# Downloading Data

Created: October 22, 2008; Updated: August 12, 2013.

## Aspera Connect

**What is Aspera software, where to get it?**

The dbGaP Authorized Access System uses Aspera, a high-speed file transfer system, to facilitate client download. It requires Aspera Connect to be installed on client's download machine. Aspera Connect is an install-on-demand browser plugin. It is available for free on the Aspera website. From the software download page, please make sure to select and install Aspera Connect instead of any other Aspera client products. Aspera Connect is available for Linux, Mac, and Windows platforms. In addition to the web user interface, Aspera Connect also includes a command line ASCP executable utility. (**04/21/2015**)

## Download Procedure

**Download using prefetch command-line utility with the cart file or SRA accession**

The principal investigator (PI) of the project or downloaders designated by the PI can download the data as soon as the data access request is approved. The recommended way of downloading dbGaP data is using the "prefetch" utility available in the NCBI SRA toolkit.

The prefetch utility can download dbGaP non-SRA and SRA data files in bulk when a cart file is provided as an argument. It can also download the data of individual SRA run when individual SRR accession is provided as an argument. The documentation of prefetch can be found from here.

The following are main steps of downloading with prefetch.

1. Download and install Aspera Connect (see here for more information).
2. Select and save data files information in a cart (.kart) file

   (For SRA data download, in addition to bulk download with cart file, the prefetch can also run with individual SRA accession, which is often preferred method for program/script directed automatic download. See the section 5 for more about this.)

   - Login to the dbGaP Authorized Access System using the eRA account login credentials. (Intramural NIH scientists and staff need their NIH email username and password).
   - Click on "My Requests" tab. The list of Approved Requests is under "Approved" sub-tab.
   - Find the table row of approved dataset, click on the link named "Request Files" in the "Actions" column.
   - On the "Access Request" page, different types of data files available for download are shown separately under different sub-tabs. To download non-SRA data, go to the "Phenotype and Genotype files" sub-tab and click on the "dbGaP File Selector" link. To download SRA data, go to the "SRA data (reads and reference alignments)" sub-tab and click on the "SRA RUN Selector" link.
   - Wait until the page loading is complete. Click on the "Help" icon on top of the page to see instruction/information about the selector).
   - Add/remove files using the facets listed in the left panel facet manager. From the right panel file list, select/unselect files by checking/unchecking checkboxes in front of the file names.
   - Once the files are selected (checked), click on the "Cart File" button (on the upper part of the page) and save the cart file (.kart).

3. Download and decrypt dbGaP data files
   - Download the **latest** version of the NCBI SRA Toolkit. Untar or unzip downloaded toolkit file.
   - Follow this dbGaP Download Guide to download dbGaP data using the sratoolkit.
4. Specific steps and commands

   Before running the download commands below, make sure the dbGaP repository key (.ngc) and the cart files are ready.
   - Download a fresh dbGaP repository key (.ngc) file and re-config the toolkit with the command below.
     *$ /path-to-your-sratoolkit-installation-dir/bin/vdb-config -i*
   - From the sratoolkit GUI interface, import the repository key
   - Download dbGaP data files
   - Run the command below to download the files specified in the cart file.
     *$ /path-to-your-sratoolkit-installation-dir/bin/prefetch --ngc /path-to-ngc-file-dir/xxxxx.ngc /path-to-your-cart-file/xxxxx.krt*
     Please make sure the sratoolkit, ngc, and cart files are on the same disk drive.
   - Decrypt downloaded files
   - The downloaded dbGaP non-SRA files need to be decrypted before use. Run the command below to decrypt the files.
     *$ /path-to-your-sratoolkit-installation-dir/bin/vdb-decrypt --ngc /path-to-ngc-file-dir/xxxxx.ngc / path-to-top-level-download-dir/*
5. Compatibility issue with older versions of sratoolkit

   If 2.9.6 or older version of the sratoolkit had been installed and used on the machine, before running above commands, the old toolkit settings need to be disabled by renaming the settings file as below.
   *$ cd ~/.ncbi*
   *$ mv user-settings.mkfg user-settings.mkfg.old*

**(12/08/2020)**

## How to Add Downloaders to Projects?

**I am a principal investigator (PI). Is it possible to allow my lab staff or collaborator to download data without sharing my eRA login credentials?**

Here is a video related to this topic. Recently improved user-interface of the dbGaP Authorized Access System allows principal investigator (PI) to designate one or more downloaders within PI's institution. A Downloader is an individual assigned by the PI to perform the time-consuming task of retrieving large data files. The downloaders can login to the dbGaP system through their own account and make download. The download is limited to the data sets approved to access and specified for downloader by primary PI.

The following is how to assign downloaders to approved datasets within all or specific projects:

1. Login to the dbGaP Authorized Access System as a PI using the eRA login credentials; If respective project hasn't yet been created, create the project and follow multiple steps to complete and submit the online application.
2. Navigate to "Downloader" page through "Downloaders" tab. Search for the name of intended downloader by the first name and last name using the search boxes.

**Note**: A downloader needs to have a valid NIH eRA Commons account or a NIH email account, and have successfully logged into the dbGaP Authorized Access System at least once. Downloader's eRA account does not need to have a PI role, but it does need to be affiliated with PI's institution.

1. Confirm to make sure the resulting user name is correct; Click on the name; select all or a specific project from the pull-down manual, and finally click on "Set downloader" button to make the assignment. The downloader's name and the projects accessible to the downloader will be displayed on the page.
2. The PI can use the "X" buttons in "Remove Role" column of downloader table to remove any downloaders or downloader's projects.

(**07/13/2011**)

## How to Become a Downloader?

**I am a data analyst working for a principal investigator (PI) who has multiple approved data access requests. How can I download PI's datasets without logging into his account?**

Here is a video related to this topic. Downloader has to be designated by the PI through the dbGaP system. Please see here for more details. Prior to be chosen as a downloader, the individual must

1. Have a valid NIH eRA Commons account affiliated with the same organization as the PI, or has an NIH email account. The eRA account does not need to have a PI role.
2. Have already completed at least one successful login to the dbGaP Authorized Access System.

(**07/12/2011**)

## Download Procedure for Downloader

**I am a downloader designated by the principal investigator (PI). How do I make download?**

The download procedure is nearly the same for PI and for downloaders. Please see here for more details. (**06/30/2011**)

## Expired Download Package

**My download package is expired. What can I do with it?**

In most of cases, the expiration interval of a download package is set to two months. You can always delete expired package and order a new one if you need to download the same data again. The new download package can include some or all of the previously downloaded files. Please see here for more details. (**06/30/2011**)

## FTP Site Availability for Downloads

**Can I use FTP instead of Aspera to download dbGaP data? I don't have large file to download.**

No, the FTP interface is no longer available for downloading dbGaP data. The Aspera Connect is the only choice. (**06/21/2011**)

# Decrypting and Extracting Data

Created: October 22, 2008; Updated: August 12, 2013.

## File Decryption

**Are downloaded files encrypted? If so, do I need to decrypt them and how?**

The following instructions are nearly identical in all supported platforms.

1. **Different treatment of SRA and non-SRA data**

   The data files distributed through the dbGaP are all encrypted by NCBI's data encryption algorithm. These files have a file suffix ".ncbi_enc", indicating that they are NCBI encrypted files. Not all encrypted data however need to be decrypted.

   The SRA (short-read-archive) data distributed through the dbGaP are encrypted **but there is no need to decrypt them**. The NCBI SRA toolkit can work directly on encrypted SRA data without decryption. Decrypted SRA data is in a binary format that is not human readable and can only be processed by the SRA toolkit anyway.

   You need NCBI SRA toolkit to work on SRA data. The SRA toolkit is a collection of utilities that can dump, extract, and convert SRA data to different data formats. The vdb-decrypt utility included in the SRA toolkit can be used to decrypt any encrypted dbGaP data.

   The dbGaP data other than SRA (non-SRA data) need to be decrypted before use. If you are only working on non-SRA data, you can download the NCBI Decryption Tool, which is a sub-set of the SRA Toolkit. It only includes utilities related to data decryption. If you already have SRA toolkit setup, you don't need to download NCBI decryption tool because the vdb-decrypt utility is included.

   Both NCBI SRA Toolkit and NCBI Decryption Tool are available from here.

2. **The dbGaP repository key**

   dbGaP repository key is a dbGaP project wide security token required for configuring NCBI SRA toolkit and decryption tools. The key is provided in a file with suffix ".ngc". It can be obtained from two places in PI's dbGaP account.

   1. The first place is the project page under "My Projects" tab, through a link named "get dbGaP repository key" in the "Actions" column. The key downloaded from here is valid to all downloaded data under the project.
   2. The second place is the download page under "Downloads" tab, through a link named "get dbGaP repository key in the "Actions" column.

3. **Toolkit Configuration and import repository key**

   The NCBI decryption tool is a subset of the SRA Toolkit. The steps of setting up both tools are nearly identical. In either case, a dbGaP repository key for the respective dbGaP project should be downloaded from PI's dbGaP account, and the tool should be first configured using "vdb-config", a command line utility available under the "bin" directory of the toolkit. See here for detailed instruction.

4. **Decrypting Non-SRA Data**

   The Non-SRA data distributed through the dbGaP need to be decrypted before used for anything. The tool named "vdb-decrypt" under NCBI sra-toolkit or NCBI decryption Tools is for data decryption.

To decrypt non-SRA data, go to the dbGaP project directory (workspace) setup through the toolkit configuration, issue the following command from a command line: It is important to remember that the command line has to be run directly from the dbGaP project directory.

A typical vdb-decrypt command should be like this:
*$ /path-to-your-sratoolkit-installation-dir/bin/vdb-decrypt --ngc /path-to-ngc-file-dir/xxxxx.ngc /path-to-top-level-download-dir/*

5. **More about NCBI SRA Toolkit**

   Please refer to the documentation of sra-toolkit for more about various utilities available under the sra-toolkit.

(**12/09/2020**)

# SRA to BAM format conversion

**We would like to get the data in BAM format but they are only available in SRA format. What can we do?**

Most of the sequencing data available through the dbGaP are in SRA format. The SRA data can be converted to BAM format using the sam-dump combined with samtools. The sam-dump utility is available under the SRA toolkit. More information about the sam-dump is available at here, and the information about the samtools can be found from here.

(**12/24/2013**)

# SRA fastq-dump Utility

**How to convert downloaded SRA data into FASTQ format?**

Please visit the section related to the fastq-dump utility in SRA Download Guide. If you have further questions regarding SRA (Short-Read-Archive) data, please directly contact NCBI's SRA group (sra@ncbi.nlm.nih.gov). They are better able to help with SRA related issues.

(**10/19/2011**)

# Data Sample and Subject ID Mapping

Created: October 22, 2008; Updated: August 12, 2013.

## 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)

# General Information regarding dbGaP Data Access

Created: October 21, 2008; Updated: December 7, 2011.

This section of the dbGaP FAQ Archive contains general information about accessing data in dbGaP that range from determining where the data you are looking for is located and the definition of "Embargo Release Dates", to the need for institutional affiliation to access data and how to manage your eRA passwords.

**To begin searching this section of the dbGaP FAQ Archive, you can either**:

- Enter your search word(s) text in the text box at the top of the page and click on the "Go" button,

**OR**

- Click on any of the "General Information regarding dbGaP Data Access" sub-categories listed in the navigation box on the right side of the page to navigate to the sub-category of your choice.

# Data Sharing Policy

Created: October 21, 2008; Updated: December 7, 2011.

**What data sharing policy one should follow when using the dbGaP data?**

NIH Policy for Sharing of Data Obtained in NIH supported or conducted Genomic Data Sharing Policy (GDS Policy) also referred to as the "GDS Policy" is published on the NIH GDS website. General questions about the GDS policy should be addressed to gds@mail.nih.gov.

(**08/22/2014**)

# Informed Consent for Genomic Research

Created: October 21, 2008; Updated: December 7, 2011.

**Would you provide some model informed consent language for studies which have a dbGaP reporting requirement?**

The dbGaP data distribution is governed by the NIH GDS policy. There is some information about informed consent for genomic research on NHGRI website that may be useful to you. The following is how to get there:

From the NHGRI website, click on the "Issues in Genetics" tab and select "Informed Consent for Genomic Research". The specific link that may be relevant is to this page.

(**08/22/2014**)

# Citing dbGaP in a Publication

Created: October 21, 2008; Updated: December 7, 2011.

**How do I cite dbGaP in the "Reference" section of a publication I am writing?**

When you write a manuscript and report data that

1. are based on the use of dbGaP study data, accessed and/or downloaded from the dbGaP web site (public pages for general and summary information or Authorized Access pages for subject-level data), please cite the USE of dbGaP study data in your manuscript by referencing the study accession (phs#), and the data use, and provide a link to the appropriate study page if at all possible, using a format like this:

   The data/analyses presented in the current publication are based on the use of study data downloaded from the dbGaP web site, under phsxxxxxx.vx.px (e.g. phs000001.v1.p1/https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000001.v3.p1);
2. are based on data that you uploaded to the dbGaP database, please reference the data upload using a format like this:

   The data/analyses presented in the current publication have been deposited in and are available from the dbGaP database under dbGaP accession phsxxxxxx.vx.px (e.g. phs001233.v1.p1).

When you have used or deposited data of more than one dbGaP study, please reference each dbGaP study separately, e.g.:

- The data/analyses presented in the current publication are based on the use of study data downloaded from the dbGaP web site, under dbGaP accession e.g. phs000001.v1.p1, phs000002.v1.p1, phs000003.v1.p1;
- The data/analyses presented in the current publication have been deposited in and are available from the dbGaP database under dbGaP accession e.g. phs001233.v1.p1, phs001234.v1.p1.

When you are citing the dbGaP database in general, without reference to specific studies, please use a format like this: dbGaP/database of Genotypes and Phenotypes/ National Center for Biotechnology Information, National Library of Medicine (NCBI/NLM)/https://www.ncbi.nlm.nih.gov/gap

When appropriate, please add dbGaP references, e.g.:

The NCBI dbGaP Database of Genotypes and Phenotypes. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L,

Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST.

Nat Genet. 2007 Oct; 39(10):1181-6.

NCBI's Database of Genotypes and Phenotypes: dbGaP.

Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M, Feolo M.

Nucleic Acids Res. 2014 Jan; 42 (Database issue): D975-9.

(01/06/17)

# Is the Data I'm looking for Authorized (Controlled) or Public Access?

Created: October 21, 2008; Updated: December 7, 2011.

## Questionnaires

**How do I access the questionnaire for the Bipolar disorder whole genome association study? Do I need to apply for authorized access?**

The questionnaires for the Bipolar disorder whole genome association study are public access documents, and you therefore do not need to apply for authorized access.

There are a number of different questionnaires associated with the Bipolar Disorder study. Use the following steps to find them:

1. Go to dbGaP's study page for the Whole Genome Association Study of Bipolar Disorder.
2. When you arrive at the study page, select the "Documents" tab from the series of tabs located under the study title, which will take you to the study's document page.
3. Look for the grey "Associated Documents" box located to the right. If you navigate through the folder structure in the box you will find all of the different documents for this study. (**HINT**: click on the file name rather than on the file icon).
4. There are a number of questionnaires under the "Cases" folder, and there is a single questionnaire under the "Controls" folder.

These documents can also be downloaded from the dbGaP GAIN Bipolar Disorder public FTP site. You can find the link to this FTP site on the study home page: under the "Access to Publicly Available Data (Public ftp)" section of the page.

As the documents on the FTP site are organized by their accession numbers only, in order to retrieve the documents you want from the FTP site, you will have to note down the document's accession number as you review the document in the document section of the study page. The accession number is located in the "Document Name and Accession" section (the first section under the title) of a document's web description. The accession number always starts with "phd".

Please note that in the case of large questionnaires, you will have to download a series of xml files and their related images. (**03/13/08**)

# Summary-Level Data

Created: October 21, 2008; Updated: December 7, 2011.

## Is Authorized Access Required for Viewing Summary-Level data?

**Where can I access aggregate level data?**

Due to the findings in the Homer et al, aggregate level data are no longer available through the dbGaP public ftp site. The data are only available via the dbGaP Authorized Access System together with the individual level data. NIH has recently implemented a new process for accessing certain aggregate datasets designated as general research use (GRU) through a single request for the dbGaP study phs000501 (Compilation of Aggregate Genomic Data Study), the data requests of which will be reviewed by the Central Data Access Committee (CDAC).

For additional information about the new process and how to request access to the compilation of aggregate data, please see here. (**07/31/2012**)

## Embargo Release Dates

### Definition

**Some dbGaP studies have an embargo date attached to them. Could you explain what this means, and if these dates apply to summary data?**

Prior to an embargo release date there is a restriction on the permission of public discussion — including manuscripts, posters, abstracts and presentations — of the data, which allows the data provider time to publish their findings.

Summary level allele frequencies and phenotype distributions are not subject to embargo release dates, however, summary level analyses, such as the provisional QC analyses ARE subject to embargo release dates. (**10/30/07**)

### Access to Data prior to Embargo Release Date

**Is it possible to get access to datasets prior to an embargo release date?**

Many dbGaP datasets are available prior to the embargo release date. When authorized to do so, you may download and analyze the controlled access data prior to the embargo release date; however, you are not permitted to publicly discuss your findings as described in the signed Data Use Certification (DUC) that is part of the application process. (**07/31/12**)

# Is Institutional Affiliation Required for Data Access?

Created: October 21, 2008; Updated: December 7, 2011.

**I want to use individual-level data to validate the code of a testing tool. Do I need institutional affiliation to access this data? Can I access non-human data instead?**

You must have an institutional affiliation and research credentials to request individual level human data from dbGaP. (**08/22/08**)

# Searching dbGaP

Created: October 22, 2008; Updated: March 23, 2009.

This section of the dbGaP FAQ Archive contains general information about where you might find specific information in dbGaP, and how you might search dbGaP for specific information.

**To begin searching this section of the dbGaP FAQ Archive, you can either**:

- Enter your search word(s) text in the text box at the top of the page and click on the "Go" button,

**OR**

- Click on any of the "Searching dbGaP" sub-categories listed in the navigation box on the right side of the page to navigate to the sub-category of your choice.

# Is the Material I Need in Public Access or Authorized (Controlled) Access?

Created: October 22, 2008; Updated: March 23, 2009.

## Questionnaires

**How do I access the questionnaire for the Bipolar disorder whole genome association study? Do I need to apply for authorized access?**

The questionnaires for the Bipolar disorder whole genome association study are public access documents, and you therefore do not need to apply for authorized access.

There are a number of different questionnaires associated with the Bipolar Disorder study. Use the following steps to find them:

1. Go to dbGaP's study page for the Whole Genome Association Study of Bipolar Disorder.
2. When you arrive at the study page, select the "Documents" tab from the series of tabs located under the study title, which will take you to the study's document page.
3. Look for the grey "Associated Documents" box located to the right. If you navigate through the folder structure in the box you will find all of the different documents for this study. (**HINT**: click on the file name rather than on the file icon).
4. There are a number of questionnaires under the "Cases" folder, and there is a single questionnaire under the "Controls" folder.

These documents can also be downloaded from the dbGaP GAIN Bipolar Disorder public FTP site. You can find the link to this FTP site on the study home page: under the "Access to Publicly Available Data (Public ftp)" section of the page.

As the documents on the FTP site are organized by their accession numbers only, in order to retrieve the documents you want from the FTP site, you will have to note down the document's accession number as you review the document in the document section of the study page. The accession number is located in the "Document Name and Accession" section (the first section under the title) of a document's web description. The accession number always starts with "phd".

Please note that in the case of large questionnaires, you will have to download a series of xml files and their related images.(**03/13/08**)

# How to Search for Specific Information in a Particular Study

Created: October 22, 2008; Updated: March 23, 2009.

## Finding Variables Represented in a Specific Study.

**How do I find out if the Framingham SHARe data includes alcohol use and smoking as variables?**

You can query the dbGaP system directly to see a short list of the variables in which you are interested. Below are the steps for an example query to get you started:

1. Go to the NEI Age-Related Eye Disease Study (AREDS) , and type "smoking" (do not use quotation marks) into the "Search Within This Study" box, located on the right side of the page, and press "go".
2. The results of this search is shown under a series of tabs labeled "Studies", "Variables", "Study Documents", and "Analyses". The number shown in parentheses next to the tab label is the number of items matching your query.
3. You can use the AND, OR, and NOT operators to make more complex queries to the system (e.g. smoking AND alcohol). (**08/14/08**)

# Submitting to dbGaP

Created: October 21, 2008; Updated: April 4, 2014.

This section of the dbGaP FAQ Archive contains general information about the dbGaP submission process.

**To begin searching this section of the dbGaP FAQ Archive, you can either**:

- Enter your search word(s) text in the text box at the top of the page and click on the "Go" button,

**OR**

- Click on any of the "Submitting to dbGaP" sub-categories listed in the navigation box on the right side of the page to navigate to the sub-category of your choice.

# Beginning the Submission Process

Created: October 21, 2008; Updated: April 4, 2014.

### The dbGaP study registration and data submission process

An outline of steps of dbGaP study registration and data submission can be found from here. (**04/04/2014**)

# NIH ICs that Support GWAS

Created: October 21, 2008; Updated: April 4, 2014.

**Could you provide a list of NIH ICs that support GWAS?**

The following NIH Institutes and Centers (ICs) support Genome Wide Association Studies (GWAS):

National Cancer Institute (NCI)

National Center for Research Resources (NCRR)

National Eye Institute (NEI)

National Human Genome Research Institute (NHGRI)

National Heart, Lung, and Blood Institute (NHLBI)

National Institute on Aging (NIA)

National Institute of Allergy and Infectious Diseases (NIAID)

National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS)

National Institute of Biomedical Imaging and Bioengineering (NIBIB)

National Institute of Child Health and Human Development (NICHD)

National Institute on Drug Abuse (NIDA)

National Institute on Deafness and Other Communication Disorders (NIDCD)

National Institute of Dental and Craniofacial Research (NIDCR)

National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)

National Institute of Environmental Health Sciences (NIEHS)

National Institute of General Medical Sciences (NIGMS)

National Institute of Mental Health (NIMH)

National Institute on Minority Health and Health Disparities (NIMHD)

National Institute of Neurological Disorders and Stroke (NINDS)

National Institute of Nursing Research (NINR)

(**11/17/2011**)