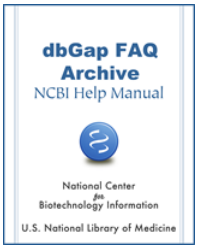
U.S. National Library of Medicine
National Center for Biotechnology Information

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



Downloading Data

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. In either case, before running the command, the sratoolkit has to be configured with the dbGaP repository key and **the command has to be issued under the dbGaP workspace (dbGaP-prj# directory**, such as dbGaP_2220) automatically created through the configuration.

The cart file currently can only be used with prefetch. Other sratoolkit utilities can work with SRR accession but not with cart file.

The prefetch can download data through the Aspera or through Http without Aspera. The path information of Aspera connect needs to be provided in the command line for Aspera download. **Aspera is more robust and faster for the downloading job with large data size.**

The documentation of prefetch can be found from here.

The following are four 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" 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. Configure SRA toolkit

   - Download the latest version of the NCBI SRA Toolkit. Untar or unzip downloaded toolkit file.
   - Follow the Protected Data Usage Guide to configure the toolkit. The toolkit needs to be configured before use, through which the dbGaP repository key will be imported.
   - The dbGaP repository key file contains the information required by the SRA Toolkit to identify approved user and the project where the downloaded dbGaP data belong. The following is how to download the dbGaP repository key.
     a. Login to PI's dbGaP account.
     b. Under the "My Projects" tab, find the project where download data belong. Click on the link named "get dbGaP repository key" in the "Actions" column. Save the key (.ngc) file.
     c. For an old download package, the dbGaP repository key of the download package can also be retrieved from the page under the "Downloads" tab. In the table row of download package, click on the link "get dbGaP repository key" in the "Actions" column, and save the key (.ngc) file.
   - The dbGaP repository key needs to be imported during the SRA toolkit configuration process. After the configuration, a dbGaP project directory, also called workspace directory, is automatically created. The default location of the dbGaP project directory is like the following.

     /home/foo/ncbi/dbGaP-2000 (on a Linux terminal) or /Users/foo/ncbi/dbGaP-2000 (on a Mac terminal) or C:\Users\foo\ncbi\dbGaP-2000 (on a Windows DOS terminal)

     Note: The location and name of the dbGaP project directory may be different if the default settings are not chosen during the configuration.

     Check and confirm the dbGaP repository directory is created. If it is created correctly, the sub-directories, such as "sra", "refseq", should be seen under it.

4. Bulk download using the "prefetch" command-line utility with cart-file

   The "prefetch" is a command line tool available in the SRA toolkit. It takes the cart file as an argument and download the selected in the cart-file.

- From a command-line terminal, go to (cd into) the dbGaP project directory created through the configuration. Issue the prefetch command as instructed in the Toolkit Documentation page. Provide the full-path location to the cart file obtained in the previous step as an argument.

  Please note that **all sratoolkit commands should be issued directly under the dbGaP project directory** (not anywhere else). From the directory, the full-path to the "prefetch" executable file under the "bin" directory of the sratoolkit should be provided as an argument. **Issuing the prefetch (or other dbGaP data related sratoolkit) command from a place other than the dbGaP project directory is the single most common error seen from many dbGaP users.**

  The following is how to run the prefetch command.

  [path-to-dbgap-project-directory] $ /path-to-sratoolkit-install-dir/bin/prefetch [options] cart-file-name

  The command first looks for the Aspera Connect installation path and carries out the download through the Aspera ascp command line utility if the path exists. If the path is not found, the command will download through Http as a failover.

  If the Aspera Connect is installed in a custom location, for large data download, it may be needed to use -a option to provide the full path to Aspera's ascp and openssh explicitly as shown below.

  prefetch -a "%ASPERA_CONNECT_DIR%/bin/ascp|%ASPERA_CONNECT_DIR%/etc/asperaweb_id_dsa.openssh" cart-file-name (or SRR accession)

  (

  For example:

  prefetch -a "/opt/aspera/bin/ascp|/opt/aspera/etc/asperaweb_id_dsa.openssh" SRR390728

)

  Where %ASPERA_CONNECT_DIR% is the full-path to the ascp connect installation directory. The default location of ascp connect installation for different platforms are the following:

  on Linux ---> /home/foo/.aspera/connect/

  on Microsoft Windows ---> C:\ProgramFiles\Aspera\Aspera Connect

  on Mac ---> /Applications/AsperaConnect.app/Contents/Resources

- Downloaded data files are saved in the "files" directory under the dbGaP project directory.

5. Prefetch with SRA accession

   - For SRA data download, in addition to downloading in bulk with the cart-file, the prefetch command can also take individual SRA accession (e.g. SRR390728) as an argument and download the data of each SRA run one by one. The command is like

     [path-to-dbgap-project-directory] $ /path-to-sratoolkit-install-dir/bin/prefetch [options] SRR#

     Please note that **all sratoolkit commands should be issued directly under the dbGaP project directory** (not anywhere else). From the directory, the full-path to the "prefetch" executable file under the "bin" directory of the sratoolkit should be provided as an argument. **Issuing the prefetch (or other dbGaP data related sratoolkit) command from a place other than the dbGaP project directory is the single most common error seen from many dbGaP users.**

     More information about the prefetch command can be found from here.

- The list of SRA accessions (SRR#) of specific data access request can be obtained from PI's dbGaP account. The following is how to find it.

  Login to primary PI's dbGaP account. Go to "My Requests" tab, find the table row of the approved data access request, and click on the "Request Files" link in the "Actions" columns. On the resulting page, go to the sub-tab named "SRA data (reads and reference alignments)". Click on the link "Get manifest (CSV format)" at the bottom of the page. The saved spreadsheet file contains the list of all SRA accessions available for the access request.

(05/03/2019)

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**)