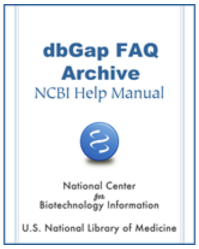
NLM Citation: GaP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2009-. Decrypting and Extracting Data.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/


dbGap FAQ Archive
NCBI Help Manual
National Center for Biotechnology Information
U.S. National Library of Medicine

Decrypting and Extracting Data

Points often Ignored When Decrypting or Extracting dbGaP Data

I used the vdb tools to work on dbGaP data but ran into errors, stating that the repository key is not found or file corrupted. What may be wrong?

The decryption and extraction of non-SRA or SRA data obtained from the dbGaP requires either NCBI Decryption Tool or SRA Toolkit available from the SRA Toolkit page of NCBI SRA website. A few important steps are involved when using the both tools. Missing one or more of these steps is the common reason that the decryption process does not go through.

1. Make sure to use the latest version of the NCBI Decryption tool or the SRA Toolkit.
2. Follow the detailed instruction in the “Configure SRA toolkit” section of the Download Procedure in an early part of this document or follow this instruction to configure the toolkit using the vdb-config, a command-line utility that can only be launched through a command-line terminal.
3. During the configuration, make sure the dbGaP repository key obtained from PI’s dbGaP account is imported.
4. After the configuration, make sure a dbGaP project directory (workspace) such as the following is created:

   /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).

   Make sure the project ID of the workspace, 2000 in above examples, maches the dbGaP project ID of downloaded data.
5. **Issue the vdb utility (such as vdb-dump or vdb-decrypt) command directly from the dbGaP-prj# directory (also called dbGaP workspace). Do not run it from any other directory.** Issuing the command from a different location is the single most common cause of problems seen from many dbGaP users.
6. The command should Include the full path to the vdb utility and the full-path to the data directory (or data file) when running the vdb utilities.

Please see more detailed instruction from here. If you have followed all steps in the instruction and the problem persists, please send detailed description and screenshots of terminal input and output to dbgap-help@ncbi.nlm.nih.gov.

(12/24/2014)

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:

```
[foo@server /home/foo/ncbi/dbGaP-2000]$ /home/foo/SRA-toolkit.version-and-platform/bin/vdb-decrypt full-path-to-top-level-directory-of-non-sra-data
```

5. **Running vdb-ump on encrypted SRA data file**

   The dbGaP SRA data downloaded using the prefetch tool (see here for more about downloading with prefetch) are located in the "files" sub-directory of the dbGaP workspace (dbGaP-prj# directory, such as dbGaP-2000). Go to the dbGaP workspace and issue vdb-dump command such as below. The command has to be run directly from the dbGaP project directory.

   Run with downloaded SRA files

   ```
   [foo@server /home/foo/ncbi/dbGaP-2000]$ /home/foo/SRA-toolkit.version-and-platform/bin/vdb-dump files/encrypted-SRR-file –T PRIMARY_ALIGNMENT -R 1 > output-file-name-with-path
   ```

6. **SRA data dump by accession without downloading sequence data**

   The dump utilities such as vdb-dump and fastq-dump can run with SRA accession (SRR run accession) and dump the data to respective format without downloading actual SRA sequence data. The command is like the following

   ```
   [foo@server /home/foo/ncbi/dbGaP-2000]$ /home/foo/SRA-toolkit.version-and-platform/bin/vdb-dump -f tab SRR#
   ```

   The command on other platforms should be very similar. Please note that the default location of dbGaP project directories on different platforms are slightly different. For example:

   /home/foo/ncbi/dbGaP-2220 (Linux)

   /Users/foo/ncbi/dbGaP-2220 (Mac)

   C:\Users\foo\ncbi\dbGaP-2220 (Windows)

   Similarly, several other SRA toolkit commands, such as fastq-dump, can be run with SRA accession in the same way. See here for more about other commands.

   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.

7. **More about NCBI SRA Toolkit**

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

(**09/16/2015**)

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**)