Published online 1 December 2013 *Nucleic Acids Research, 2014, Vol. 42, Database issue* ***D975–D979***
doi:10.1093/nar/gkt1211

NCBI's Database of Genotypes and Phenotypes: dbGaP

Kimberly A. Tryka*, Luning Hao*, Anne Sturcke, Yumi Jin, Zhen Y. Wang, Lora Ziyabari, Moira Lee, Natalia Popova, Nataliya Sharopova, Masato Kimura and Michael Feolo*

Information Engineering Branch, National Center for Biotechnology Information, Bethesda, MD 20894, USA

Received October 9, 2013; Revised and Accepted November 4, 2013

ABSTRACT

The Database of Genotypes and Phenotypes (dbGap, http://www.ncbi.nlm.nih.gov/gap) is a National Institutes of Health-sponsored repository charged to archive, curate and distribute information produced by studies investigating the interaction of genotype and phenotype. Information in dbGaP is organized as a hierarchical structure and includes the accessioned objects, phenotypes (as variables and datasets), various molecular assay data (SNP and Expression Array data, Sequence and Epigenomic marks), analyses and documents. Publicly accessible metadata about submitted studies, summary level data, and documents related to studies can be accessed freely on the dbGaP website. Individual-level data are accessible via Controlled Access application to scientists across the globe.

SUMMARY

The Database of Genotypes and Phenotypes (dbGaP) (1) is a National Institutes of Health (NIH)-sponsored repository charged to archive, curate and distribute information produced by studies investigating the interaction of genotype and phenotype. It was launched in 2006 in response to the development of NIH's Genome Wide Association Study (GWAS) policy and provides unprecedented access to very large genetic and phenotypic datasets funded by NIH and other agencies world wide. Scientists from the global research community may access all public data and apply for controlled access data. Information about submitted studies, summary level data and documents related to studies can be accessed freely on the dbGaP website (http://www.ncbi.nlm.nih.gov/gap). Individual-level data can only be accessed after a Controlled Access application, stating research objectives and demonstrating the ability to adequately protect the data, has been approved (https://dbgap.ncbi.nlm.nih.gov/aa).

DATA—ACCESSIONED OBJECTS

The information contained in dbGaP includes individual-level molecular and phenotype data, analysis results, medical images, general information about the study and documents that contextualize phenotypic variables, such as research protocols and questionnaires. Submitted data undergoes quality control and curation by dbGaP staff before being released to the public. Information in dbGaP is organized as a hierarchical structure and includes the accessioned objects, phenotypes (as variables and datasets), various molecular assay data (SNP and Expression Array, Sequence and Epigenomic marks), analyses and documents. Each of these will be described in their own section below.

STUDIES

The data archived and distributed by dbGaP is organized as studies. Studies may either be stand-alone or combined in a 'parent study/child study' hierarchy. Parent or 'top level' studies may have any number of child studies (also referred to as substudies). However, study hierarchy is limited to two levels (parent and child only). In other words, substudies may not have substudies. Studies, whether parent or child, can contain all types of data ascertained in genetic, clinical or epidemiological research projects such as phenotype and molecular assay

*To whom correspondence should be addressed. Tel: 1 301 402 2874; Fax: 1 301 480 2918; Email: feolo@ncbi.nlm.nih.gov
Correspondence may also be addressed to Kimberly A. Tryka. Tel: +1 301 402 5119; Fax: +1 301 480 2918; Email: trykak@ncbi.nlm.nih.gov
Correspondence may also be addressed to Luning Hao. Tel: +1 301 443 5926; Fax: +1 301 480 2918; Email: hao@ncbi.nlm.nih.gov
Present Address:
Michael Feolo, Information Engineering Branch, National Center for Biotechnology Information, Bethesda, MD 20894, USA.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

information that are linked via subject and sample IDs. Studies also often contain documents, such as questionnaires and protocols, which help to contextualize the phenotype and genotype data. Study data are distributed by consent groups each of which contain all data from a set of study participants who have signed the same consent agreement; therefore the data delivered for a single consent group will all have the same Data Use Limitations (DULs) for future research use. Studies are assigned a unique accession number.version (e.g. phs000001.v3.p1) which should be used when citing the study.

While the data found in studies can vary widely based on both the number of participants and the variety of deposited phenotypic and molecular data, all studies include basic descriptive metadata such as study title, study description, inclusion/exclusion criteria, study history, disease terms, publications related to the study, names and affiliations of the principal investigators (PIs), and sources of funding. This information, as well as the research statements of all approved users, is publicly available on the study's report page.

DATASETS AND VARIABLES

Phenotypic data values are submitted to dbGaP as tabular files or datasets. The dbGaP phenotype variables consist of two parts: the data values and the description of the data in the accompanying data dictionary. Each cell (value) in a dataset is stored in a relational database and is mapped to a phenotype variable and subject. Phenotype variable metadata are provided by the submitter via a data dictionary for each dataset. Discrepancies between the data and data dictionaries are reconciled by dbGaP curators in consultation with submitters. Variables are grouped into datasets. A dataset's version will change when a variable inside the dataset is added, updated, or deleted. Variables, and when appropriate, datasets are linked to sections of documents when possible. Individual-level phenotype data are only available through the dbGaP Authorized Access System. Public summary-level variable information is available on the dbGaP website and ftp site.

GENOTYPE DATA

Genotype data hosted at dbGaP comprise individual-level genotypes and aggregated summaries, both of which are distributed exclusively through the dbGaP Authorized Access System. The types of data available include DNA variations, SNP assay, DNA methylation (epigenomics), copy number variation, as well as genomic/exomic sequencing. RNA data types such as expression array, RNA seq and eQTL results are also available. For details about the accepted format of submitted genotype files please see the dbGaP submission guide. Genotype data files are compressed and archived into tar files for distribution. The files are explicitly named to indicate file content, raw data (cel and idat), genotype calls (genotype) and locus annotations (marker info). Genotype calls are usually clustered according to file format and genotyping platform, including one sample per file (indfmt), multiple-sample matrix (matrixfmt) and pre-defined variance call format (vcf) or in other popular formats. They will be accompanied by a sample-info file for subject lookup and consent status. The consent code and consent abbreviation are also embedded in the file name.

ANALYSES

Analyses are assigned accessions using the general dbGaP format pha######.v (e.g. pha003690.1). As a result of the large volume of data generated and concerns regarding participant confidentiality, many genetic epidemiological analysis results have not been published. But because individual-level data is only accessible through Controlled Access, dbGaP can archive, integrate and distribute full association results. Additionally, after removing identifiable elements, like counts and frequencies, analysis results are displayed in the public dbGaP browser which dynamically links to NCBI annotation resources, like dbSNP, Gene and RefSeq. These public views can be found through the 'Analyses' link on the study page and they can be downloaded from the ftp site for use or display in other browsers. The full original submitted analyses are fully accessible through the dbGaP Authorized Access System.

DOCUMENTS

The dbGaP encourages investigators to submit documents related to their studies, such as protocols, patient questionnaires, survey instruments and consent forms, along with their data. These documents provide valuable information and context for subsequent researchers who will apply for and download datasets. All submitted documents are available publicly and can be used by anyone interested in gaining a better understanding of the phenotypic data found in a study. Each document is accessioned using the general dbGaP format phd######.v#. Documents submitted to dbGaP are converted and marked up using common XML format. Converting documents into XML allows all documents to be treated uniformly in the database (aiding indexing and discovery) and to be displayed in a single HTML style. Additionally, the XML format allows annotation which is used to create live links between the documents and other portions of the dbGaP website, such as variable report pages.

The XML used by dbGaP for document processing is an extension of NLM's Archiving and Interchange Tag Set Version 2.3 (http://dtd.nlm.nih.gov/archiving/2.3/). A copy of the extension is publicly available at http://dtd.nlm.nih.gov/gap/2.0/wga-study2.dtd, and documentation for the extension is located at http://dtd.nlm.nih.gov/gap/2.0/doc/wga-document.html.

SUBMITTING DATA

The NIH strongly supports the broad sharing of de-identified data generated by NIH-funded investigators and facilitates data sharing for meritorious studies that

are not NIH-funded. Decisions about whether non-NIH-funded data should be accepted are made by individual NIH Institutes and Centers (IC); ICs will not accept data unless its submission is compatible with NIH's GWAS policy.

NIH-FUNDED STUDIES

Institutional certification, as well as basic information about the study, is required when submitting data to dbGaP.

- 'Institutional certification' consists of a letter signed by the PI and an institutional official that confirms permission to submit data to dbGaP. NIH has developed Points to Consider for IRBs and Institutions to assist institutions in their review and certification of an investigator's plan for submission of data to dbGaP.
- 'Basic information' consists of items like the title of the study, a description and history of the study, inclusion and exclusion criteria, listing of study investigators and funding information.

PIs should familiarize themselves with the 'NIH Points to Consider' document which provides information about: the NIH GWAS Data Sharing Policy; benefits of broad sharing of data through a central data repository; risks associated with the submission and subsequent sharing of such data; safeguards designed to protect the confidentiality of research participants; and specific points for institutional review boards to consider during review and certification of PIs' data-submission plans. The PIs must contact their NIH program official (PO) to begin the submission process.

NON-NIH-FUNDED STUDIES

To submit non-NIH-funded data to dbGaP the following information will need to be provided:

- 'Institutional certification' as described in the last section. To provide this, someone from the institution or organization will need to be registered in eRA Commons. Information regarding registration is available from the eRA Commons website. (Note: the review of a PI's request can be initiated without the certification, but the review process will be expedited if the GWAS staff receives the certification at time of submission.)
- 'Basic information' about the study, as described in the previous section.
- 'The NIH IC' that most closely aligns with the research. A list of ICs can be found at http://www.nih.gov/icd/.
- 'Whether the study has been published or accepted for publication.' If it has, the PI should provide documentation (i.e. the publication citation or a copy of any correspondence indicating that an article about the study has been accepted for publication).

The PI should submit all information and the certification to gwas@mail.nih.gov. Once the GWAS staff receives the documents, they will forward them to the appropriate IC program administrator for consideration. The IC program administrator will contact the PI with any questions and/or to notify the PI of the IC's decision. The PI is encouraged to consult with the Program Officer/Director (PO/PD) and/or IC GWAS Program Administrator (GPA) at an NIH IC to discuss the project, data sharing plan and data certification to complete the registration process. Non-NIH-funded projects should contact dbGaP Help.

SEARCHING dbGaP

All publicly released dbGaP studies can be queried from the search box on the top of the dbGaP homepage. Queries can be very simple, just a keyword of interest ('cancer') or complex, making use of search fields and Boolean operators ('cholesterol[variable] AND phs000001'). More complex searches can be facilitated by using the 'Advanced Search' which helps create queries via a web form. Additionally, complex queries can contain Boolean operators. For example: Cancer[Disease] AND True[Study Has SRA Components], returns a list of all studies having SRA data and where the PI has assigned the keyword 'cancer' as a disease term. As with all other NCBI resources, the searches in dbGaP are performed using the Entrez search and retrieval system. Please see the Entrez chapter of the NCBI handbook for general guidance on forming Entrez queries.

Once a search query is executed and results returned, clicking on an item's name or accession will lead to a page listing more specific information about that object. This information is of particular importance to those users who want to find out more about a study before deciding whether or not to apply for Authorized Access. (Note that on each of the different pages, one can examine other objects in the study by using the navigational aid along the right-hand edge of the page.)

VARIABLES ON THE PUBLIC WEBSITE

Phenotype variables can either be found by doing a search on the dbGaP home page and then linking to individual variable pages, or by choosing the 'Variables' tab when looking at the website of a study. When viewing the 'Variables' tab the phenotypes are generally grouped into broad categories for ease of browsing. These grouping can be found to the right-hand side of any variable report web page.

DOCUMENTS ON THE PUBLIC WEBSITE

There are multiple pathways to documents through the dbGaP website. On the dbGaP home page, the newest studies are listed under the 'Latest Studies' heading, with the most direct route to documents being the orange 'D' icons. A gray icon means there are no documents associated with the study. Beneath that section, the 'List Top Level Studies' link leads to a searchable listing of all

studies and documents, with an advanced search option available for building document-specific queries. On a study page, clicking the Documents tab will open the study's default document, with a folder tree on the right to explore the rest, and a 'Search Within This Study' box that will search document text. Variable pages may also link to documents in which they are annotated, through the 'See document part in context' links. Documents for each study are also available on the dbGaP ftp site as a downloadable zip file, which includes the PDFs, XML and images. The ftp site is accessible from study pages by clicking the link under 'Publicly Available Data'.

dbGaP ftp SITE

The ftp site includes a directory for every study, which contains a directory for every version of a study, as well as a directory where analyses are found. Currently, each version of a study contains directories for documents, phenotype variable summaries, manifests and release notes when applicable. Manifests describe the files available in each consent category while release notes describe the history of the released files as well as giving details of any changes made from previous versions.

The variable summaries and the data dictionaries are delivered as XML files with an accompanying XSL file which produces the HTML rendering of the file that can be viewed on a browser. The documents directory contains at least one .zip file that holds the XML files, images and PDF versions of the documents in a study. In cases where there are a large number of documents the files may be separated into separate.zip files for xml, images and PDFs.

dbGaP AUTHORIZED ACCESS

Data distribution by dbGaP is governed by the NIH's policies and procedures for managing GWAS data. Information related to these policies can be found on the NIH GWAS website. Questions related to GWAS policy can be directed to gwas@mail.nih.gov.

The individual-level data is only available to authorized users. Requests for data and data downloads are managed through the dbGaP Authorized Access System (dbGaP-AA), a web platform that handles request submission; manages reviewing and approval processes carried out by signing officials (SOs) and Data Access Committees (DACs); and facilitates secured high-speed large data download for approved users.

The dbGaP data are distributed by consent groups. That is, the data are grouped by subjects that have agreed to the same set of DULs. The data can only be selected by consent group when making data access requests. There are no overlapping subjects between multiple consent groups within a study. The data requests are also reviewed and approved by consent group. Therefore it is very important that requesters understand the DULs of consent groups before they apply for dbGaP data access. Access requests must include a brief research use statement that explains how the research objectives conform to the DULs or the requested data. These research use statements are listed on the public study page to provide transparency to the public about the research being conducted with individual-level data.

Each data file distributed through the dbGaP has an embargo release date. The data access policy requires that the results obtained from analyzing the dbGaP data are not published before the embargo release date. To access the Authorized Access system, non-NIH users must have an NIH eRA Commons account with a PI role.

RELATED RESOURCES

Phenotype–genotype integrator

The Phenotype-Genotype Integrator (PheGenI) (2) merges NHGRI GWAS catalog data with several databases housed at the National Center for Biotechnology Information (NCBI), including Gene, dbGaP, OMIM, GTEx and dbSNP. This phenotype-oriented resource, intended for clinicians and epidemiologists interested in following up results from GWAS, can facilitate prioritization of variants to further pursue, study design considerations and generation of biological hypotheses. Users can search based on chromosomal location, gene, SNP or phenotype and then view and download results. These include annotated tables of SNPs, genes and association results, and a dynamic genomic sequence viewer. PheGenI is under active development and currently the phenotype search terms are based on MeSH and will be enhanced with additional options such as SNOMED and HPO in the future. Please note that PheGenI does not display all *P*-values from each dbGaP hosted analysis. Specifically, only *P*-values $<10^4$, and/or the lowest 100 *P*-values are included for each analysis. Public view of all analysis results can be seen using the dbGaP analysis and clicking on the genome browser; for full analysis and aggregate statistics such as allele frequencies, please apply for controlled access.

The dbGaP genome browser

The genome-wide association results hosted at the dbGaP are displayed through the dbGaP genome browser, where they can be viewed along the human genome.

The dbGaP genome browser can be accessed through the analysis page of a given dbGaP study. For example, under the 'Analyses' tab of the dbGaP study phs000585.v1.p1. If there are multiple analyses, selection can be made from the right panel. The link named View association results in Genome Browser leads to the chromosomal viewer and each region (block) there contains results from all tested loci within. The color is coded for the smallest *P*-value in that block.

FUNDING

Intramural Research Program of the US National Institutes of Health, National Library of Medicine. Funding for open access charge: Intramural Research

Program of the National Institutes of Health, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Mailman,M.D., Feolo,M., Jin,Y., Kimura,M., Tryka,K., Bagoutdinov,R., Hao,L., Kiang,A., Paschall,J., Phan,L. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet*., **39**, 1181–1186.
2. Ramos,E.M., Hoffman,D., Junkins,H.A., Maglott,D., Phan,L., Sherry,S.T., Feolo,M. and Hindorff,L.A. (2013) Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet*., **22**, 144–147.