Scope and Access
The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype. Such studies include genome-wide association studies (GWAS), medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits. The advent of high-throughput, cost-effective methods for genotyping and sequencing has provided powerful tools that allow for the generation of the massive amount of genotypic data required to make these analyses possible.

dbGaP provides two levels of access. Open access through the dbGaP homepage (https://www.ncbi.nlm.nih.gov/gap) provides the public with summaries of studies, the contents of measured variables as well as original study document text. Access to individual-level data, including phenotypic data tables and genotypes, requires varying levels of authorization. Information on controlled access is available at: https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login. Access requires an eRA Commons login.

dbGaP Homepage
The homepage of dbGaP (shown below) provides a central entry point to access the data from this database. Entering terms in the search box and clicking the Search button (A) performs a search against this database. The Access dbGaP Data section (B) lists links to browse and access public content from the database or apply for controlled access to get the detailed data. Important Links and help (C) provides dbGaP-specific documentation and help. The Latest Studies (D) presents a selective list of GWAS dataset recently deposited to dbGaP with their titles linking to corresponding entries, and additional summary information given in the columns to the right. Integrated searches for phenotype and genotype data can be done using the Phenotype-Genotype Integrator (E).

The Advanced Search link (F) provides an alternative way to find studies of interest. It offers many sets of filters to facet existing studies into various categories to allow quick identification of studies of interest. For more information, see this webinar from the NCBI YouTube channel: https://www.youtube.com/watch?v=ePQ9p2SL_wM

Legend for colored icons under the Details column: V=variable, D=documents, A=analysis, and S=SRA data.
Data Available from dbGaP and Controlled Access

Data deposited in dbGaP come from various types of GWAS. These include longitudinal, case control, and cohort studies. For phenotype, the data include information collection forms, description of phenotypes, standards of measurement, and details of the individual phenotypes. For genotype, the data includes genotype calls and their quality scores from various platforms. If sequencing and expression analyses are included, the data will be brokered through other NCBI databases such as Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA). The summary of the phenotype and genotype data, associated through appropriate statistical methods, plus the analysis details are also available. Two levels of data access, public and controlled, are adopted to protect the privacy of study participants. Data that can potentially be used to establish personal identity of the participants are placed under restricted access. These include individual phenotypes, genotypes, sequence reads, expression profiles, epigenetic markers and full result sets. Information on the data collection, standards of phenotype observation and measurements, platforms for genotype determination, and the final summary of association analyses are available to the public through the dbGaP homepage.

For validated research needs with institutional support, a principal investigator can apply for controlled access to a specific dataset using the page shown to the right. Once a request is granted, the necessary information and credentials will be provided to the applicant so the specific dataset can be downloaded for further analysis. Further assistance is available by contacting dbGaP help: dbgap-help@ncbi.nlm.nih.gov

Searching in dbGaP

Data available for public access can be searched and retrieved through the dbGaP homepage. In the example below, the disease “macular degeneration” is combined with filter “[has analysis]” to retrieve related studies with full analysis (A).

This is indicated by the purple-colored icon (B), which links directly to the analysis summary. In this study, the genotype analyses were conducted on different platforms so multiple results are retrieved and only one is shown here. The “Browse genome for …” link (C) points to an informative graphical display of SNPs analyzed for the study on this platform (p. 3). Clicking the title of the study (D) displays the text-based information for that study. Functions provided by the “Advanced” page (E) can be used to construct a more complex query to retrieve studies satisfying more specific criteria, through the usage of field-limited query terms and history number. An email
Graphic Summary Through “Browse Genome for …”

“Browse genome for …” is linked to a graphical summary of the analysis results (shown below). Here, chromosomes are divided into bins of fixed sizes (A). These regions are color-coded according to the significance of the association between the identified genotype and the phenotype being studied, with bins in red indicating strong association (B). Clicking a bin opens a more detailed GaP Browser display for the region with a gene level resolution (C) or higher. In that display, the region shown is marked by the coordinates (D), which is further divided into smaller bins of fixed-length. A summary of sub-regions with significant phenotype-associated genotypes and number of SNPs found is displayed in the GWAS catalog track (E), while results from a specific analysis is shown in a track below (F), with SNPs for the region displayed in a scatter plot just below. Hovering-over a bin highlights the SNPs present in the region (G). The genome annotation pane (H), located under and aligned to the GWAS track, highlights the gene features found for the region. Clicking a bin in the track under the GWAS (F) zooms in to a much more detailed display for both panes (not shown). Displays can also be adjusted using controls at the top and in the left sidebar. Hovering over a control will display the online help.

Other Data in dbGaP

The title of a study, those listed on the dbGaP homepage (A) or in dbGaP search results (B), links to documents available under that study. This display groups the information available for the study into different categories and places them under different tabs.

The “Study” tab (C) provides a general description on the goal of the study. The variables measured are listed under the “Variables” tab (D). Study documents with detailed background information and rationale for conducting the study are under the “Documents” tab (E). A summary of the analysis result with link to genomic display is given under the “Analysis” tab (F). The summary of available datasets provided under the “Datasets” tab (G), with the list of molecular data summed up in the “Molecular Data” tab (H).