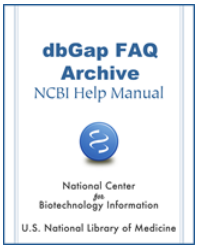
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: GaP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2009-. Individual-level Data: General Questions. 2008 Oct 21 [Updated 2009 Mar 23].
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



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**)