

DOWNLOADING HRS GENETIC DATA FROM DBGAP

TUTORIAL SUMMARY

- Overview of HRS genetic data resources
- dbGaP
- Downloading the right data
- Final steps



Public Data

- Polygenic Score Data (PGS)

Sensitive Data

- Telomere Data

dbGaP approval,
HRS download

- Candidate Gene
- Exome files

dbGaP approval,
dbGaP download

- Genotype data

DBGAP

WHAT IS DBGAP?

- “The Database of Genotypes and Phenotypes (**dbGaP**) 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”
- <https://www.ncbi.nlm.nih.gov/gap>
- dbGaP is part of the National Center for Biotechnology Information suite of resources

WHAT IS DBGAP?

Tryka, Hao, Sturcke et al. (2013). The Database of Genotypes and Phenotypes (dbGaP) and PheGenI. The NCBI Handbook, 2nd edition, Bethesda (MD): National Center for Biotechnology Information (US)
Available from:
<https://www.ncbi.nlm.nih.gov/books/NBK154410/>

WHAT IS DBGAP?

Health and Retirement Study

The screenshot shows the NCBI dbGaP website interface. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' links. Below this is a search bar with 'dbGaP' selected in a dropdown menu and 'Health and Retirement Study' entered in the search field. A red arrow points to the search bar. To the right of the search bar is a 'Search' button and a 'Help' link. Below the search bar is a banner for 'dbGaP' with a description: 'The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.' Below the banner are three columns of links: 'Access dbGaP Data', 'Resources', and 'Important Links'. At the bottom, there is a 'Latest Studies' section with a table of study results.

NCBI Resources How To

dbGaP dbGaP Health and Retirement Study Search Help

dbGaP

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.

Access dbGaP Data

- [Advanced Search](#)
- [Controlled Access Data](#)
- [Public FTP Download](#)
- [Collections](#)
- [Summary Statistics](#)

Resources

- [dbGaP Data Browser](#)
- [Phenotype-Genotype Integrator](#)
- [dbGaP RSS Feed](#)
- [Software](#)
- [dbGaP Tutorial](#)

Important Links

- [How to Submit](#)
- [FAQ](#)
- [Code of Conduct](#)
- [Security Procedures](#)
- [Contact Us](#)

Latest Studies

| Study | Embargo Release | Details | Participants | Type Of Study | Links | Platform |
|---|---------------------|---|--------------|---------------|-------|-----------------------------------|
| phs001242.v1.p1 Small genomic insertions form enhancers that misregulate oncogenes | Version 1: | V D A S | 1 | Sequencing | | MiSeq |
| | Version 1.2: passed | | | | | SureSelect Human All Exon v.2 Kit |



- Must have an eraCommons account
- PI status or download permission on an IRB approved project

- NCBI has excellent YouTube videos on each of these steps

This module:

- Finding the right data
- Downloading
- Final steps

DOWNLOADING THE RIGHT DATA

ONCE YOU HAVE APPROVED ACCESS

| | | | | | | |
|--|------------------------------|--|-------|--------------|-----------------------|--------------|
| phs000428.v1.p1 Health and Retirement Study (HRS) | Version 1: passed embargo | | 12507 | Longitudinal | Links | HumanOmni2.5 |
| phs000428.v2.p2 Health and Retirement Study (HRS) | Versions 1-2: passed embargo | | 15620 | Longitudinal | Links | HumanOmni2.5 |



- [Browse/Search](#)
- Authorized Access**
- [Help](#)

Logged in as **USER** | [Log out](#)

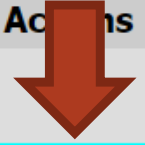


- [Beacon](#)
- [Data Browser](#)
- [My Projects](#)
- My Requests**
- [Downloads](#)
- [My Profile](#)

Request List

Approved (1)

| # | Study, Consent | Status | Expiration | Access |
|---------|--|---------------------|------------|---|
| PI: | , Project # : | | | |
| 35171-5 | Genetics Resource with the Health and Retirement Study (phs000428.v2.p2) Non-profit research use only (phs000428.v2.p2.c1), JAAMH | Data access GRANTED | 2017-05-05 | Request Files Processing History |



| Available Phenotype and Genotype Files | 1653 Gb |
|---|---------|
| Health and Retirement Study (HRS) (phs000428.v2.p2.c1) | 1653 Gb |
| StudyMeta Files | 276 Kb |
| Phenotype Files | 690 Kb |
| Genotype Files | 1653 Gb |
| phg000207.v2.CIDR_HRS_phase12.genotype-calls-indfmt.MULTI.HapMap.tar | 25 Gb |
| phg000207.v2.CIDR_HRS_phase12.genotype-calls-indfmt.c1.NPR.set1.tar | 389 Gb |
| phg000207.v2.CIDR_HRS_phase12.genotype-calls-indfmt.c1.NPR.set2.tar | 382 Gb |
| phg000207.v2.CIDR_HRS_phase12.genotype-calls-indfmt.c1.NPR.set3.tar | 20 Gb |
| phg000207.v2.CIDR_HRS_phase12.genotype-calls-matrixfmt.c1.NPR.tar | 2867 Mb |
| phg000207.v2.CIDR_HRS_phase12.genotype-qc.MULTI.tar | 517 Mb |
| phg000207.v2.CIDR_HRS_phase12.marker-info.MULTI.tar | 362 Mb |
| phg000207.v2.CIDR_HRS_phase12.raw-data-idat.MULTI.HapMap.tar | 12 Gb |
| phg000207.v2.CIDR_HRS_phase12.raw-data-idat.c1.NPR.set1.tar | 186 Gb |
| phg000207.v2.CIDR_HRS_phase12.raw-data-idat.c1.NPR.set2.tar | 183 Gb |
| phg000207.v2.CIDR_HRS_phase12.raw-data-idat.c1.NPR.set3.tar | 10 Gb |
| phg000207.v2.CIDR_HRS_phase12.sample-info.MULTI.tar | 6830 Kb |
| phg000515.v1.HRS_phase123_imputation.genotype-imputed-data.c1.NPR.tar | 123 Gb |
| phg000515.v1.HRS_phase123_imputation.genotype-qc.MULTI.tar | 960 Kb |
| phg000515.v1.HRS_phase123_imputation.sample-info.MULTI.tar | 4010 Kb |
| phg000841.v1.CIDR_HRS_phase3.genotype-calls-indfmt.MULTI.HapMap.tar | 4205 Mb |
| phg000841.v1.CIDR_HRS_phase3.genotype-calls-indfmt.c1.NPR.set1.tar | 190 Gb |
| phg000841.v1.CIDR_HRS_phase3.genotype-calls-indfmt.c1.NPR.set2.tar | 4385 Mb |
| phg000841.v1.CIDR_HRS_phase3.genotype-calls-matrixfmt.c1.NPR.tar | 715 Mb |
| phg000841.v1.CIDR_HRS_phase3.genotype-qc.MULTI.tar | 398 Mb |
| phg000841.v1.CIDR_HRS_phase3.marker-info.MULTI.tar | 496 Mb |
| phg000841.v1.CIDR_HRS_phase3.raw-data-idat.MULTI.HapMap.tar | 2230 Mb |
| phg000841.v1.CIDR_HRS_phase3.raw-data-idat.c1.NPR.set1.tar | 101 Gb |
| phg000841.v1.CIDR_HRS_phase3.raw-data-idat.c1.NPR.set2.tar | 2316 Mb |
| phg000841.v1.CIDR_HRS_phase3.sample-info.MULTI.tar | 1710 Kb |
| phg000842.v1.CIDR_HRS_phase123.genotype-calls-matrixfmt.c1.NPR.tar | 9946 Mb |
| phg000842.v1.CIDR_HRS_phase123.genotype-qc.MULTI.tar | 327 Mb |
| phg000842.v1.CIDR_HRS_phase123.sample-info.MULTI.tar | 280 Kb |

Imputed data

Genotype data

Create download request

HRS

Data-request #56084

- What is aspera and how to download and install it?
- How should I decrypt data?
- How can I change transfer speed?



[AsperaConnect](#) browser plugin must be installed before you can download files directly from your browser.

Command line (ascp) download

- Use [dbgap-aspera-download.pl](#) perl script that internally calls *ascp*, but hides some specific usage details;
OR
- Run *ascp* manually using command line below.

```
"%ASPERA_CONNECT_DIR%\bin\ascp" -QTr -l 300M -k 1 -i  
"%ASPERA_CONNECT_DIR%\etc\asperaweb_id_dsa.openssh" -W
```

- This command includes a download ticket which will be valid for a limited time. Once the ticket is expired you will need to visit this page again to get the command with an updated ticket.
- %ASPERA_CONNECT_DIR%** – path to Aspera Connect directory. You may have to adjust some ascp-specific command line keys for your environment and network connectivity. Please refer to the [AsperaConnect user manual](#) for more information.

ONCE THE DATA ARE DOWNLOADED

PROCESS

- **Download** (using FASP protocol from Aspera) ✓
- **Decrypt** (using SRA toolkit from NCBI Decryption Tools) ✓
- **Untar/Unzip** (using untar and unzip commands) ✓
- **Analyze!** (using standard genetic analysis programs, e.g. PLINK) ✓

This ends the web tutorial providing a general overview of the HRS dbGaP download and content

This tutorial was produced at the University of Michigan with funding from the National Institute on Aging.

Comments and questions may be sent to hrsquestions@umich.edu