Accessing GTEx Data in dbGaP and the SRA
Taylor Young • **The Broad Institute of MIT and Harvard (GTEx LDACC)**

GTEx PROJECT COMMUNITY MEETING — JUNE 18
www.broadinstitute.org/gtexmtg
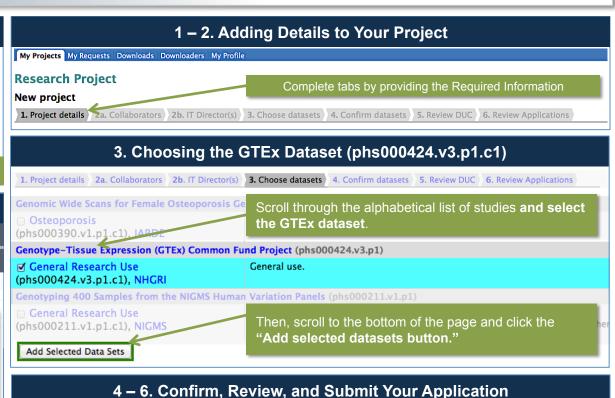
BROAD INSTITUTE

Required Information

1. Research Use Statement
2. Name of institution's signing official and IT director
3. Decryption password (save this!)
4. List of collaborators (all must have accounts at eRA Commons)
5. GTEx accession number: **phs000424.v3.p1.c1**

GO TO: www.ncbi.nlm.nih.gov/gap

Completing the Application

dbGaP

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the re... that have investigated the interaction of genotype and phenotype.

Click here

Access dbGaP Data

Apply for Controlled Access Data

Public Data via ftp Download

Association Results Browser

Phenotype-Genotype Integrator

Click "Apply for Controlled Access Data" on the dbGaP homepage. Log in and follow instructions to create a new project.

1 – 2. Adding Details to Your Project

My Projects | My Requests | Downloads | Downloaders | My Profile

Research Project

New project

1. Project details | 2a. Collaborators | 2b. IT Director(s) | 3. Choose datasets | 4. Confirm datasets | 5. Review DUC | 6. Review Applications

Complete tabs by providing the Required Information

3. Choosing the GTEx Dataset (phs000424.v3.p1.c1)

1. Project details | 2a. Collaborators | 2b. IT Director(s) | 3. Choose datasets | 4. Confirm datasets | 5. Review DUC | 6. Review Applications

Genomic Wide Scans for Female Osteoporosis Ge...

☐ Osteoporosis (phs000390.v1.p1.c1), JARDE

Genotype–Tissue Expression (GTEx) Common Fund Project (phs000424.v3.p1)

☑ **General Research Use** (phs000424.v3.p1.c1), NHGRI — General use.

Genotyping 400 Samples from the NIGMS Human Variation Panels (phs000211.v1.p1)

☐ General Research Use (phs000211.v1.p1.c1), NIGMS

[Add Selected Data Sets]

Scroll through the alphabetical list of studies **and select the GTEx dataset**.

Then, scroll to the bottom of the page and click the **"Add selected datasets button."**

4 – 6. Confirm, Review, and Submit Your Application

1. Project details | 2a. Collaborators | 2b. IT Director(s) | 3. Choose datasets | 4. Confirm datasets | 5. Review DUC | 6. Review Applications

Wait for approval (~2 weeks). Once approved, log in and download your data (see Slides 2 and 3).

Downloading Data from dbGaP

Required Materials:
1. Approved application for controlled access
2. Username and password of PI or designated downloader
3. Decryption password (specified in the application)
4. Aspera software (www.asperasoft.com)

GO TO: www.ncbi.nlm.nih.gov/gap

Access dbGaP Data
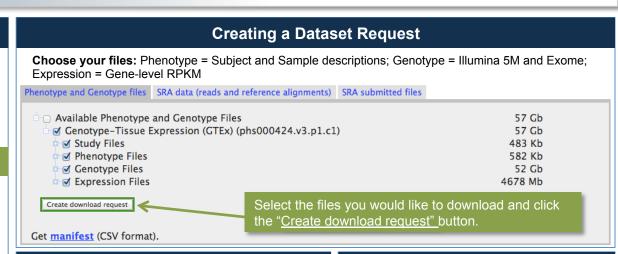
Apply for Controlled Access Data

Click here

1. Click "Apply for Controlled Access Data"
2. Log in to the authorized access system
3. Click on the "My Requests" tab

Phenotype, Genotype & Expression Data

on Fund Project (phs000424.v3.p1) l.c1), NHGRI | GRANTED | 10 | **Request Files** Processing History

Click here

Click "Request Files" to create a request for data.

Creating a Dataset Request

Choose your files: Phenotype = Subject and Sample descriptions; Genotype = Illumina 5M and Exome; Expression = Gene-level RPKM

| Phenotype and Genotype files | SRA data (reads and reference alignments) | SRA submitted files |

Available Phenotype and Genotype Files	57 Gb
Genotype-Tissue Expression (GTEx) (phs000424.v3.p1.c1)	57 Gb
Study Files	483 Kb
Phenotype Files	582 Kb
Genotype Files	52 Gb
Expression Files	4678 Mb

Create download request

Get **manifest** (CSV format).

Select the files you would like to download and click the "Create download request" button.

Downloading Your Datasets

dbGaP
genotypes and phenotypes

Logged in as **Taylor Young** | Log out

Browse/Search | Authorized Access | Help

Data-request #31507
- What is aspera and how to download and install it?
- How should I decrypt data?
- How can I change transfer speed?

Browser download (using AsperaConnect plugin) into:

1. new directory – you will be asked for download location.
   - if process fails – it can be continued later from this page (which can be accessed from "Downloads" tab).
2. default location – change it in Aspera preferences
   - default download directory is usually set to "Desktop", which is hardly good place for large files

Datasets can be downloaded with the **Aspera software** either through a web-browser or the 'ascp' **command line tool**.

Decrypting Your Datasets

You will also receive an email with a link to your data. This email will have a link to the **NCBI provided decryption tools** in the SRA toolkit.

2. NCBI Decryption Tools latest release binaries and md5 checksums*:
   - CentOS Linux 64 bit architecture
   - CentOS Linux 32 bit architecture
   - Ubuntu Linux 64 bit architecture
   - MacOS 64 bit architecture
   - MacOS 32 bit architecture
   - MS Windows 64 bit architecture
   - MS Windows 32 bit architecture

Downloading BAMs from the SRA

Required Materials:
1. Approved application for controlled access
2. Username and password of PI or designated downloader
3. Decryption password (specified in the application)
4. Aspera software (www.asperasoft.com)

GO TO: www.ncbi.nlm.nih.gov/gap

Access dbGaP Data
Apply for Controlled Access Data

Click here

1. Click "Apply for Controlled Access Data"
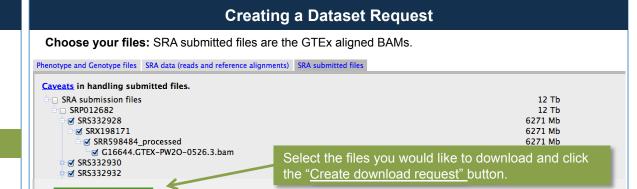2. Log in to the authorized access system
3. Click on the "My Requests" tab

Phenotype, Genotype & Expression Data

Click here

on Fund Project (phs000424.v3.p1) [ ] GRANTED 10 l.c1), NHGRI

Request Files
Processing History

Click "Request Files" to create a request for data.

Creating a Dataset Request

Choose your files: SRA submitted files are the GTEx aligned BAMs.

Phenotype and Genotype files | SRA data (reads and reference alignments) | SRA submitted files

Caveats in handling submitted files.

- SRA submission files — 12 Tb
  - SRP012682 — 12 Tb
    - SRS332928 — 6271 Mb
      - SRX198171 — 6271 Mb
        - SRR598484_processed — 6271 Mb
          - G16644.GTEX-PW2O-0526.3.bam
    - SRS332930
    - SRS332932

Create download request

Select the files you would like to download and click the "Create download request" button.

Downloading Your Datasets

dbGaP
genotypes and phenotypes
Browse/Search | Authorized Access | Help

Logged in as **Taylor Young** | Log out

Data-request #31507
- What is aspera and how to download and install it?
- How should I decrypt data?
- How can I change transfer speed?

Browser download (using AsperaConnect plugin) into:

1. new directory – you will be asked for download location.
   - if process fails – it can be continued later from this page (which can be accessed from "Downloads" tab).
2. default location – change it in Aspera preferences
   - default download directory is usually set to "Desktop", which is hardly good place for large files

Datasets can be downloaded with the **Aspera software** either through a web-browser or the 'ascp' **command line tool**.

Tips

- Download the manifest, it will indicate which files you've downloaded previously and help you find files specific to a subject or tissue site
- Mark the top level box for "SRP012682" if you want to select all files
- Use the "--overwrite=never" flag when downloading BAMs with the Aspera ascp command line tool to avoid downloading files that have been previously downloaded
- All BAMs that pass technical QC are submitted to the SRA, the Analysis Working Group defines a set of files that pass a more thorough QC