# Accessing GTEx Data in dbGaP

Taylor Young
The Broad Institute

BROAD INSTITUTE

# Contents

1. Overview of applying for access to GTEx data in dbGaP
2. Overview of downloading data from dbGaP
3. Overview of downloading data from the SRA
4. Accessing the dbGaP exchange area
   - Note: that this requires an extra step when completing your dbGaP application.
   - Note: if you’ve missed this the first time around, you can update an existing application.

# Overview of the Application

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

## Accessing GTEx Data in dbGaP and the SRA

Taylor Young • The Broad Institute of MIT and Harvard (GTEx LDACC)

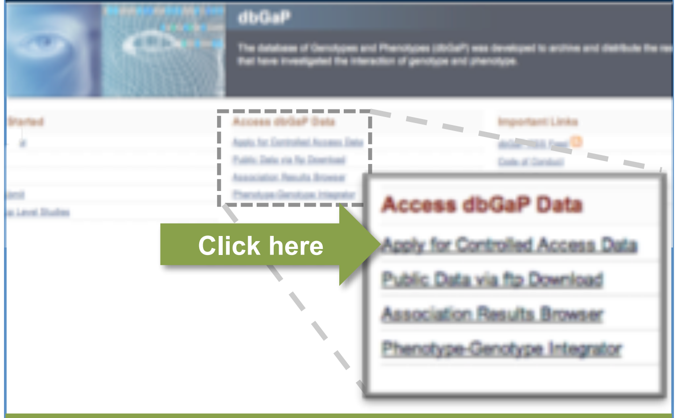
BROAD INSTITUTE

### 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**

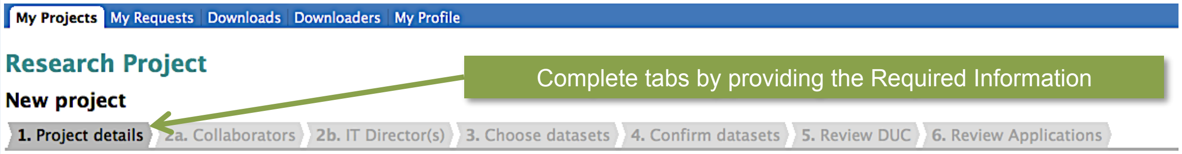
**GO TO:** www.ncbi.nlm.nih.gov/gap

### Completing the Application


dbGaP
Access dbGaP Data
Apply for Controlled Access Data
Public Data via ftp Download
Association Results Browser
Phenotype-Genotype Integrator
Important Links
Click here

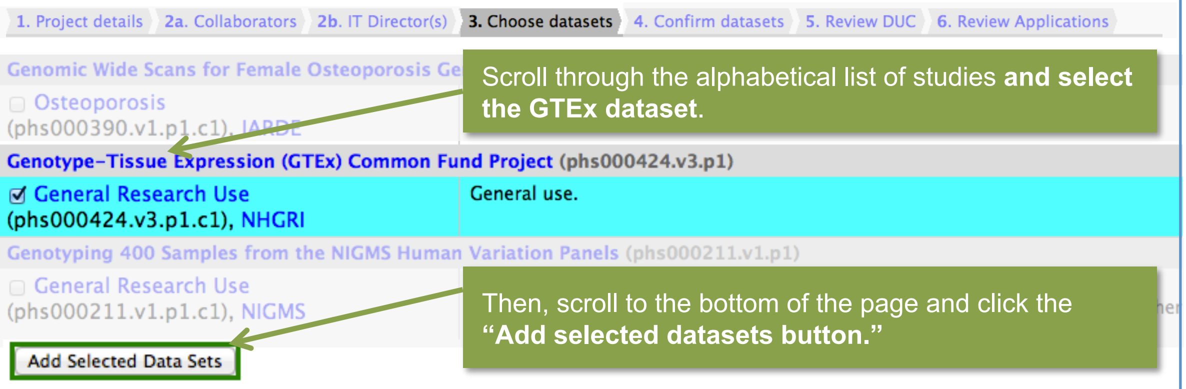
Click "Apply for Controlled Access Data" on the dbGaP homepage. Log in and follow instructions to create a new project.

### 1 – 2. Adding Details to Your Project



Complete tabs by providing the Required Information

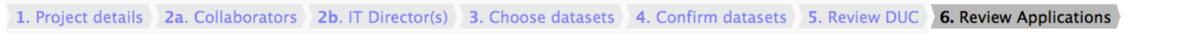
### 3. Choosing the GTEx Dataset (phs000424.v3.p1.c1)


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), IARDE
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



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

NOTE: At step 3, the GTEx study is listed about half way down the page. Checking this box provides access to publicly available data. It **DOES NOT** provide access to the exchange area.

# Overview of Data Access: dbGaP

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

## Accessing GTEx Data in dbGaP and the SRA

Taylor Young • The Broad Institute of MIT and Harvard (LDACC)

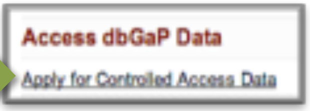
BROAD INSTITUTE

### 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

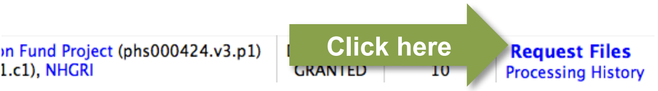

Access dbGaP Data

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

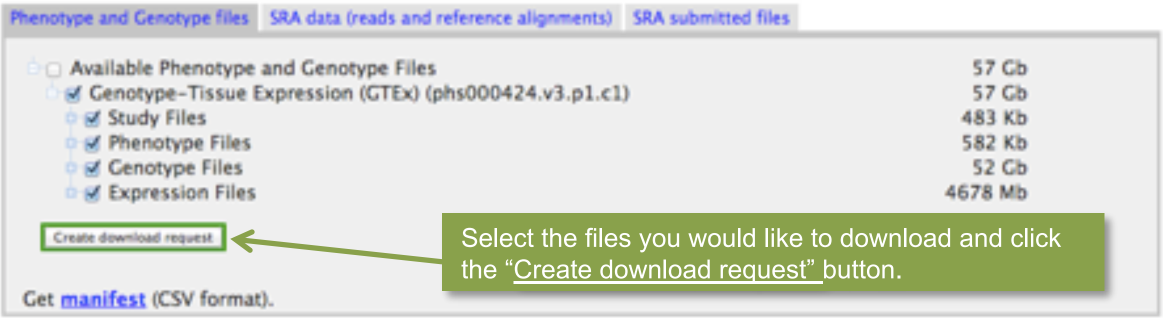

on Fund Project (phs000424.v3.p1) 1.c1), NHGRI | GRANTED | 10 | Request Files Processing History

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

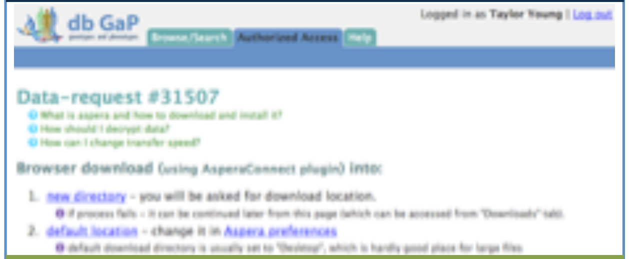

Phenotype and Genotype files
SRA data (reads and reference alignments)
SRA submitted files

| | |
|---|---|
| □ 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 |

Get manifest (CSV format).

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

### Downloading Your Datasets



Data-request #31507

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

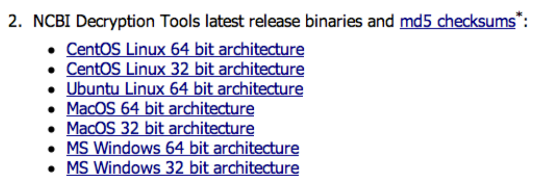
Browser download (using AsperaConnect plugin) into:

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

NOTE: BAMs need to be downloaded from the SRA, see the next page

# Overview of Data Access: SRA

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

## Accessing GTEx Data in dbGaP and the SRA

**Taylor Young • The Broad Institute of MIT and Harvard (LDACC)**

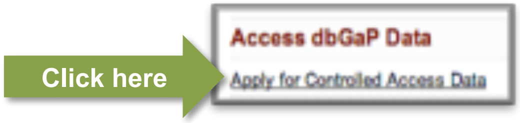
BROAD INSTITUTE

## Downloading BAMs from the SRA

**Required Materials:**

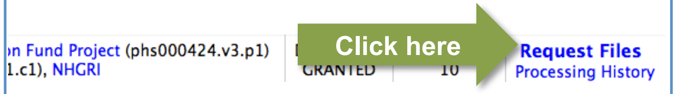
1. Approved application for controlled access
2. Username and password of PI or designated downloader
3. Decryption password (specified in the application)
4. Aspera software (www.asperasoft.com)

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



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

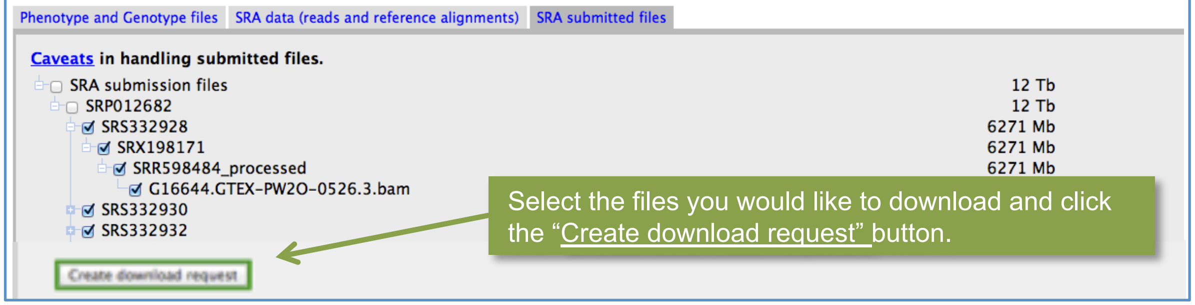


| | | | |
|---|---|---|---|
| on Fund Project (phs000424.v3.p1)<br>1.c1), NHGRI | GRANTED | 10 | Request Files<br>Processing History |

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

## Creating a Dataset Request

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

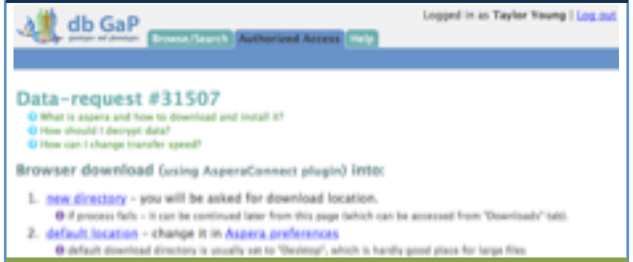

Phenotype and Genotype files
SRA data (reads and reference alignments)
SRA submitted files
Create download request

**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 | |

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

## Downloading Your Datasets



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:

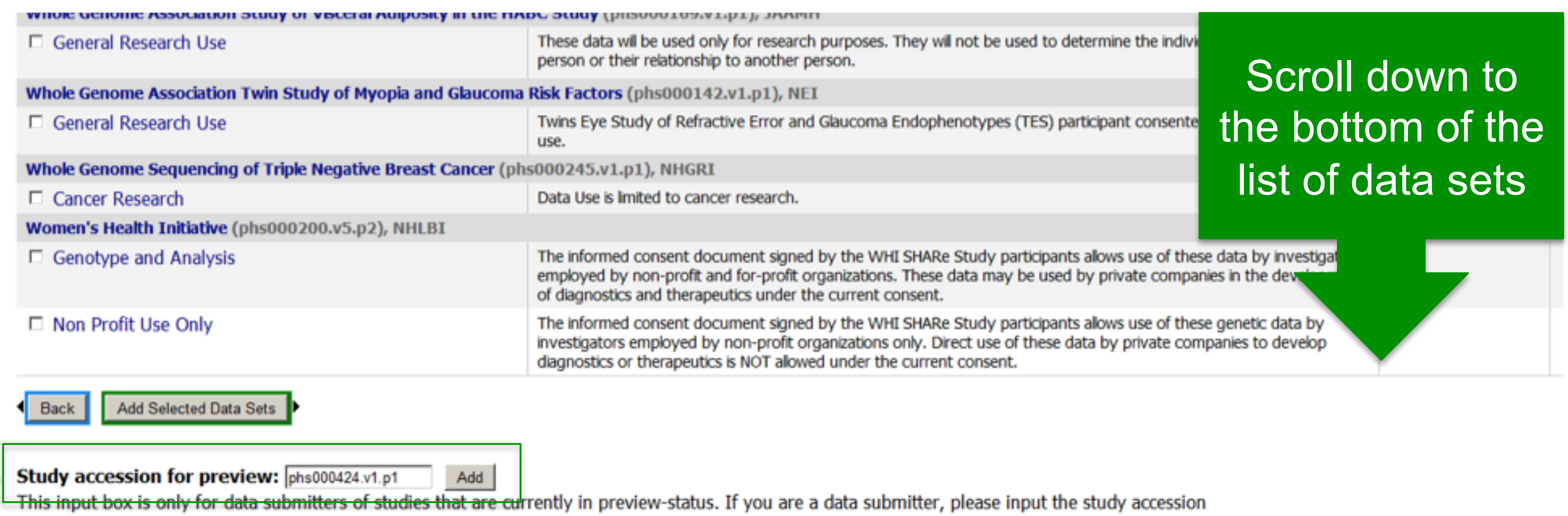
1. new directory - you will be asked for download location.
   - If process fails - it can be continued later from this page (which can be accessed from "Downloads" tab).
2. default location - change it in Aspera preferences
   - default download directory is usually set to "Desktop", which is hardly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

# Accessing the Exchange Area



| | |
|---|---|
| **Whole Genome Association Study of Visceral Adiposity in the HABC Study** (phs000169.v1.p1), JAAMH | |
| □ General Research Use | These data will be used only for research purposes. They will not be used to determine the indivi person or their relationship to another person. |
| **Whole Genome Association Twin Study of Myopia and Glaucoma Risk Factors** (phs000142.v1.p1), NEI | |
| □ General Research Use | Twins Eye Study of Refractive Error and Glaucoma Endophenotypes (TES) participant consente use. |
| **Whole Genome Sequencing of Triple Negative Breast Cancer** (phs000245.v1.p1), NHGRI | |
| □ Cancer Research | Data Use is limited to cancer research. |
| **Women's Health Initiative** (phs000200.v5.p2), NHLBI | |
| □ Genotype and Analysis | The informed consent document signed by the WHI SHARe Study participants allows use of these data by investigat employed by non-profit and for-profit organizations. These data may be used by private companies in the dev of diagnostics and therapeutics under the current consent. |
| □ Non Profit Use Only | The informed consent document signed by the WHI SHARe Study participants allows use of these genetic data by investigators employed by non-profit organizations only. Direct use of these data by private companies to develop diagnostics or therapeutics is NOT allowed under the current consent. |

**Study accession for preview:** phs000424.v1.p1 Add

This input box is only for data submitters of studies that are currently in preview-status. If you are a data submitter, please input the study accession

Scroll down to the bottom of the list of data sets

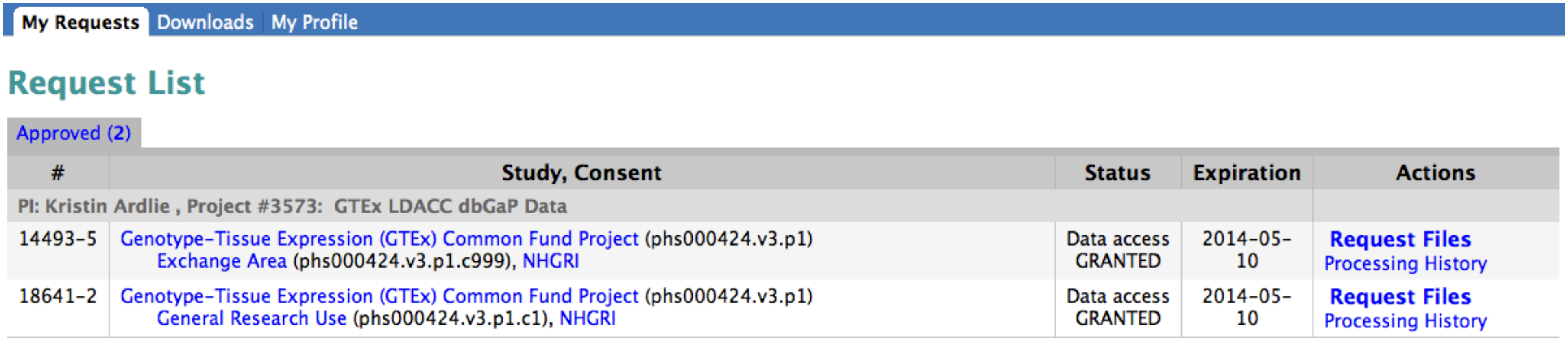
**Location in the application**: When you are selecting data sets, if you want access to the publicly available dbGaP/SRA files **AND** the exchange area, you need to add both to your application.

Access to dbGaP publicly available data is show on slide 3, follow those instructions.

To add the exchange area, when you are selecting data sets, scroll to the bottom of the list and manually enter “phs000424.v3.p1” in the box labeled “Study accession for preview.”

# Accessing the Exchange area

When your application is approved, you should see TWO entries for GTEx data:



## Request List

Approved (2)

| # | Study, Consent | Status | Expiration | Actions |
|---|---|---|---|---|
| PI: Kristin Ardlie , Project #3573:  GTEx LDACC dbGaP Data | | | | |
| 14493-5 | Genotype-Tissue Expression (GTEx) Common Fund Project (phs000424.v3.p1)<br>Exchange Area (phs000424.v3.p1.c999), NHGRI | Data access GRANTED | 2014-05-10 | **Request Files**<br>Processing History |
| 18641-2 | Genotype-Tissue Expression (GTEx) Common Fund Project (phs000424.v3.p1)<br>General Research Use (phs000424.v3.p1.c1), NHGRI | Data access GRANTED | 2014-05-10 | **Request Files**<br>Processing History |