

## EPI25 Genomic Data Sharing

**Scope:** Genomic data (Array and Whole Exome Sequencing) access via Google Buckets for approved Epi25 Research Proposals

### **Purpose of this document:**

- Outline the responsibilities of each party in the sharing of Epi25 genomic data
- Outline the process for accessing Epi25 genomic data to use in Epi25-approved analyses

### **Important Considerations:**

- It is expected that all analyses will be done in the cloud environment. Genomic data should **not** be downloaded to external institutions. All genomic data files reside in Google Buckets. Access is managed by the Stanley Center for Psychiatric Research at the Broad Institute.
- Genomic data will be made available as VCF file(s). Access may be granted to an Epi25 Master VCF including all cases and controls submitted by Epi25 sites.
- Publicly-available control datasets used in the primary Epi25 analyses are not available for use in secondary analyses *unless* the Epi25 Lead Investigator on the proposal can apply for approval via dbGaP. Please contact the Broad PM for details on the control datasets.
- VCFs are accessed by links (pointers) to the Buckets that contain them or to the VCFs directly. Access is granted as outlined in the process below. This design eliminates the need for multiple copies of a Master VCF or Sub-VCF.
- The Google bucket that contains the Master VCF has 'Requester Pays' enabled (<https://cloud.google.com/storage/docs/requester-pays>) and shared with the Account credentials supplied from the Epi25 Lead Investigator via a Google Bucket

### **Responsibilities:**

Epi25 Lead Investigator of the Research Proposal is responsible for the following:

- Submitting approved requests to the Broad Project Manager (PM) for Epi25 genomic data access
- Ensuring that data is **not** downloaded
- Establishing a Google Billing Account
- Funding the analyses

Broad is responsible for the following:

- Managing the access to all Epi25 genomic data

- Updating the genomic data files as new data are generated and released (e.g. at the end of each funding cycle when new variant callsets are ready)

## **Process**

1. Epi25 Lead Investigator submits a secondary analysis proposal to the Epi25 Strategy Committee for review and approval. For instructions on this process, navigate to the Epi25 website (<http://epi-25.org/joining-epi25/>) and scroll down to the “**To Use Epi25 Data**” section. Applications are reviewed on the monthly Strategy Committee calls.
2. Upon approval of the proposal (after review on the SC call and 2 weeks for consortium comment/questions), Epi25 Lead Investigator **contacts the Broad PM for data access**. The Epi25 Lead Investigator should determine the genomic data required for the analyses:
  - a. Which data types?
    - Whole exome sequencing
    - Whole genome genotype array (GSA-MD)
  - b. If raw data (bams for WES data or intensity files for GSA-MD data) are required, please follow up with the Broad PM to discuss options.
3. Epi25 Lead Investigator creates a Google project.
  - a. Instructions to create a Google Billing account (required to create a Google project) can be found [here](#).
4. Epi25 Lead Investigator sends the following to the Broad PM:
  - a. The approved Epi25 Research Proposal
  - b. Details of the genomic data required
  - c. Account credentials from the established Google Project that will be used to access the vcf file(s):
    - Google Project Name
    - Google project service account
  - d. If applicable, approvals for the publicly-available control datasets
5. Broad PM grants access to the appropriate Google Buckets containing the requested data.
6. Epi25 Lead Investigator accesses genomic data files in the Google bucket.

**Contact:**

Felecia Cerrato  
fcerrato@broadinstitute.org

**Appendix I. Schema of Epi25 data access**

