Data Transfer and Management

Data Storage

Workspace bucket

both in the cloud environment

Persistent Disk

Can be shared with your collaborators

Only available to you

Conduct analyses using numerous tool

Input data is localized here

Used by:

RStudio Jupyter Galaxy PROJECT

Output data is transferred here for long-term storage and sharing

gsutil

50 The cloud environment comes with 50GB of default space

Can be accessed using the terminal within the workspace

File directory is navigated through bash commands

RStudio: */home/rstudio*
Jupyter: */home/jupyter*

Pro-tip:

Anything you save outside the mounted Persistent Disk (PD) directory will be lost when the cloud environment is deleted! For more permanent storage, save your files to the workspace bucket

Getting Data into the Workspace

In your workspace, go to the DATA tab, and click on the *Import* button on the top left

Upload TSVs

1. Download sample TSV from AnVIL
2. Edit in a spreadsheet editor

   Cells can only include alphanumeric characters, "-" and "_" . No spaces are allowed.
3. Save as "tab-delimited text" or "tab-separated values"
4. Upload to workspace

Open Data Uploader

Access data uploader → Choose data collection → Upload files

You can also use the cloud bucket to upload files or entire folders!

Use **gsutil** command within your workspace:

```
gsutil cp {path/datafile}
gs://{workspacebucket}
```

You can also add a reference file to the Reference data table in your workspace, and use it as a workflow input variable by calling the file in the *Attributes* field of the workflow Input configuration form.

Managing Data with Tables

Data tables on AnVIL let you "store" and organize data files in your workspace - no matter where they are in the cloud.

Reading inputs directly from a data table allows you to:

- Process data seamlessly across all samples or the entire dataset
- Streamline analysis by chaining WDL workflows together without repeated manual setup
- Effortlessly analyze specific data subsets without repeated manual configuration.
- Integrate data from multiple sources into a single table for enhanced statistical analysis

For more information on how to manage data with tables, refer here.

Downloading data locally

Open workspace of interest → Select DATA tab → Select files of interest → Select *Export* button → Export as a TSV, to another workspace, or copy to clipboard

Stopping Compute Environment

Stop compute to free resources & minimize costs

Need data on your persistent disk?

Yes → Delete cloud environment: incur small cost for keeping the persistent disk

No → Delete both cloud environment and persistent disks: no costs incurred

Submitting Datasets into AnVIL

AnVIL can also help you host your data with these requirements:

1. All submitted human data should be based on the GRCh37 or GRCh38 human reference assemblies
2. Register data with appropriate NCBI resource
3. All individual-level human genomic and phenotypic data must conform to the NIH Genomic Data Sharing Policy

For additional information on the submission process, and data submission resources, refer here.