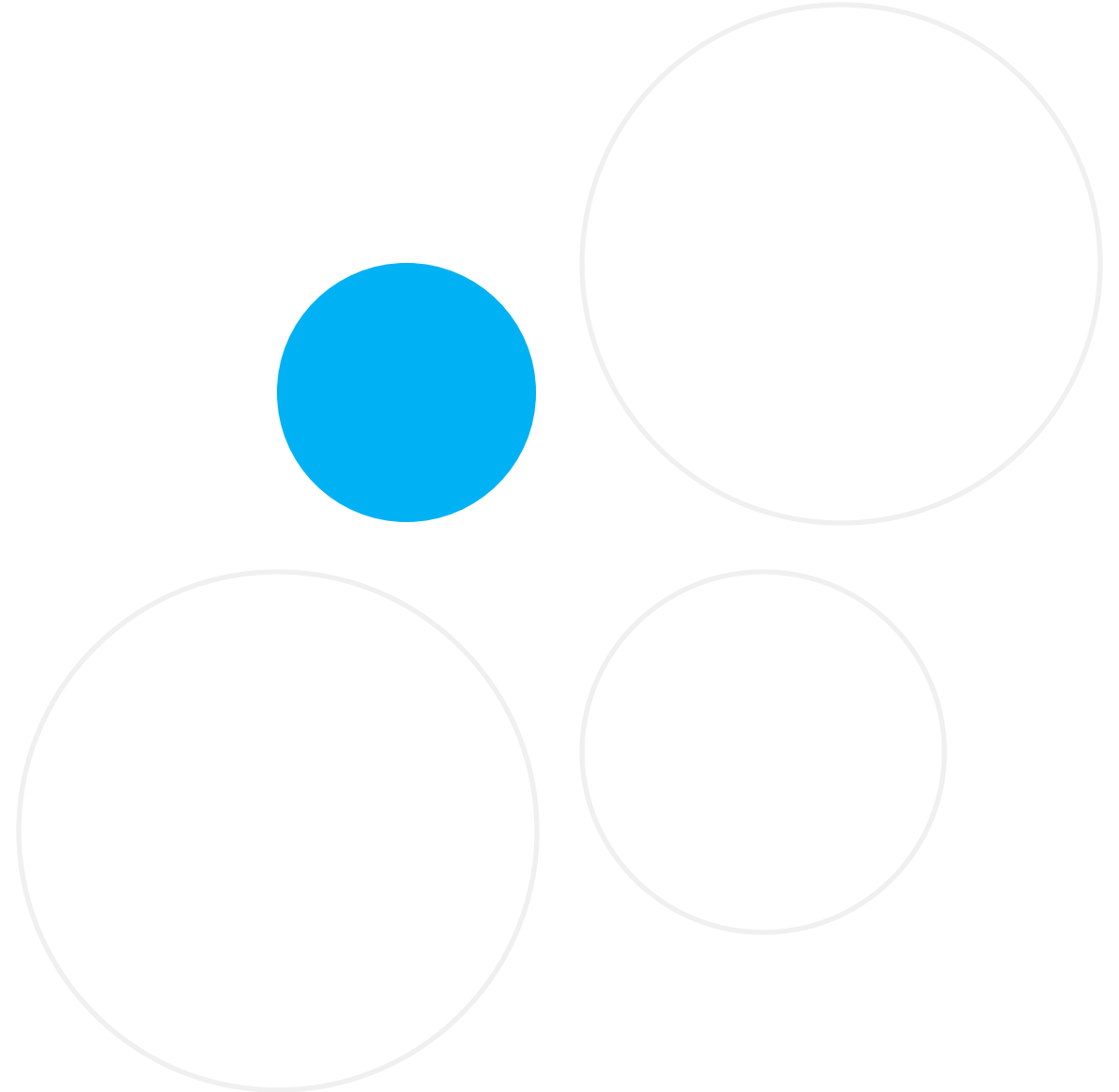


# **SC Jamboree**

## **Copying files from Synapse to Anvil workspace**

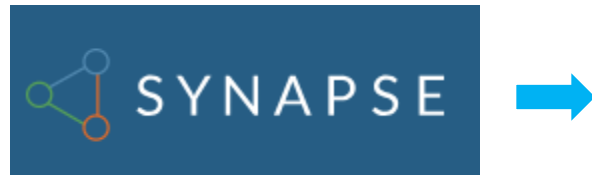
Kundaje Lab,  
Eila Arich-Landkof

---



# What?

<https://www.synapse.org/#!/Synapse:syn51669616>



gs://fc-secure-0a879173-/FASTQ/igvf\_b01\_LeftCortex/syn51669616/

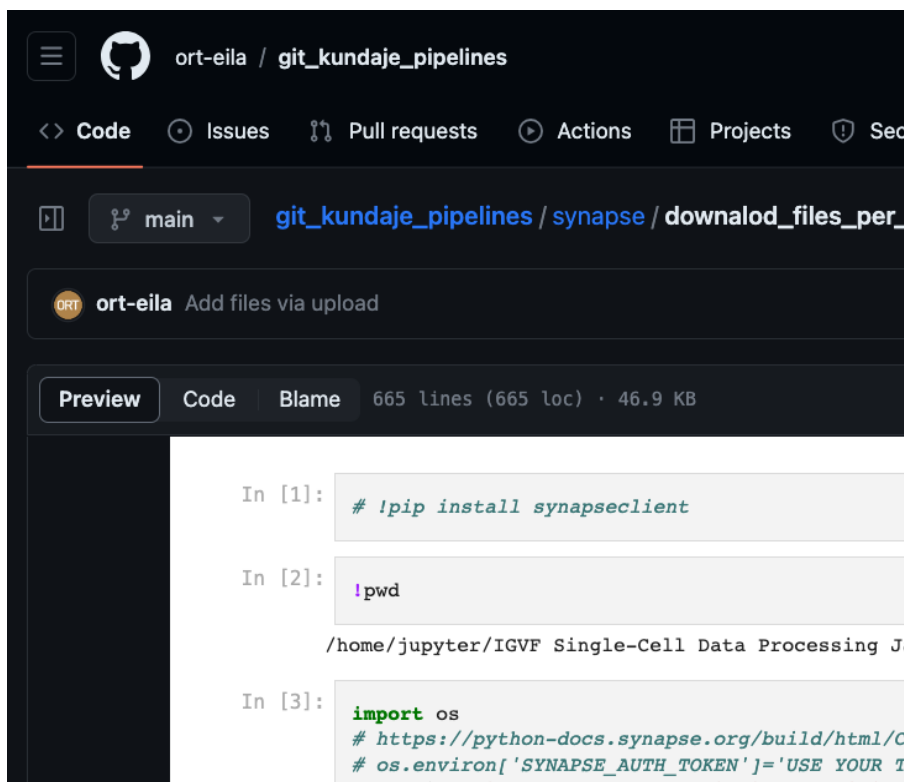


Name	
▶ Kallisto processed count matrix	
next1_B01_13A_R1.fastq.gz	
next1_B01_13A_R2.fastq.gz	
next1_B01_13B_R1.fastq.gz	
next1_B01_13B_R2.fastq.gz	
next1_B01_13C_R1.fastq.gz	
next1_B01_13C_R2.fastq.gz	
next1_B01_13D_R1.fastq.gz	

DATA	ANALYSES	WORKFLOWS	JOB HISTORY
RT DATA	Files / FASTQ / igvf_b01_LeftCortex / syn51669616		
	<input type="checkbox"/> Name		
	next1_B01_13A_R1.fastq.gz/		
	next1_B01_13A_R2.fastq.gz/		
	next1_B01_13B_R1.fastq.gz/		
	next1_B01_13B_R2.fastq.gz/		
	next1_B01_13C_R1.fastq.gz/		
	next1_B01_13C_R2.fastq.gz/		

# How? – programmatically

## [Github link](#)



The screenshot shows a GitHub repository page for 'ort-eila / git\_kundaje\_pipelines'. The repository is on the 'main' branch. The file 'download\_files\_per...' is selected. The file content is displayed in a code editor with a dark theme. The code is as follows:

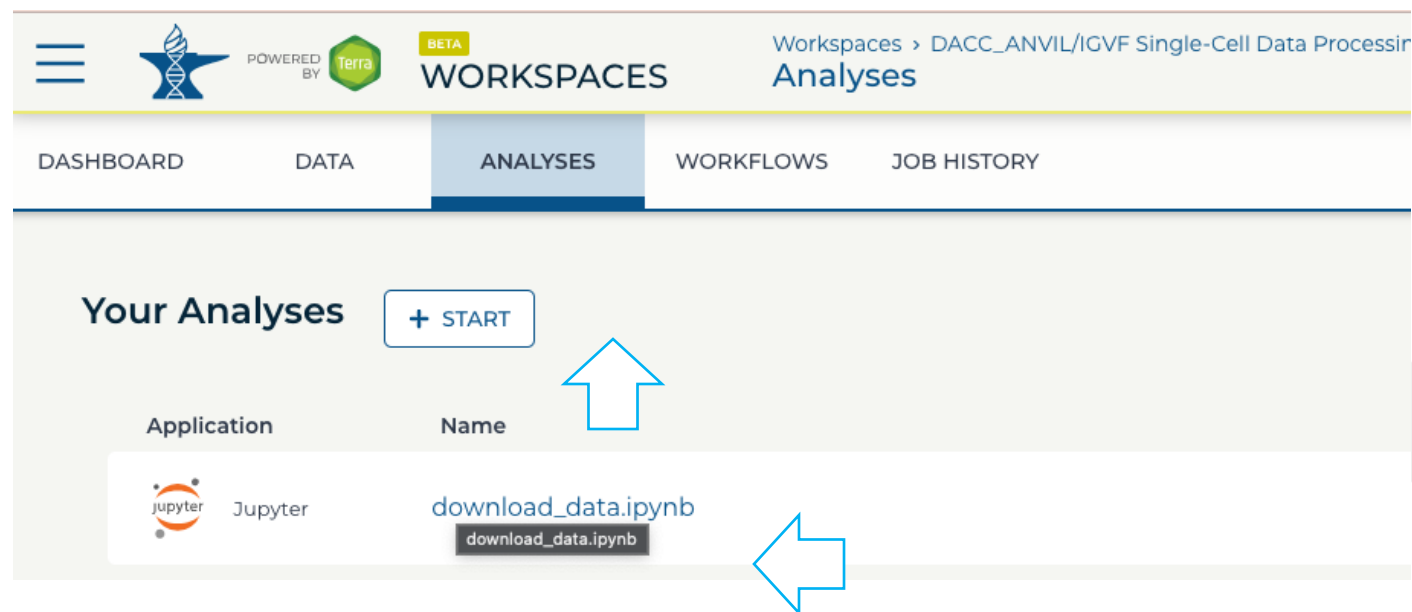
```
In [1]: # !pip install synapseclient

In [2]: !pwd

/home/jupyter/IGVF Single-Cell Data Processing J...

In [3]: import os
# https://python-docs.synapse.org/build/html/C
# os.environ['SYNAPSE_AUTH_TOKEN']='USE YOUR T
```

## Anvil workspace



- Saves time & Bugs

# Script Steps

<https://www.synapse.org/#!/Synapse:syn51669616>



gs://**bucket name**/FASTQ/**project name**/synapse ID/



1. Get the list of files from synapse based on its dataset id
  - <https://www.synapse.org/#!/Synapse:syn51669616>
2. Create the destination path for the Anvil workspace
  - gs://**bucket name**/FASTQ/**project name**/synapse ID/
  - gs://**fc-secure-0a879173-62d3-4c3a-8fc3-e35ee4248901**/FASTQ/**igvf\_b01\_LeftCortex**/syn51669616/
3. Loop on the list of files that are available on synapse dataset and copy to destination
  1. If you run it more than once, files that were downloaded, will not be downloaded again

# Destination path on Anvil

- A path on Google Cloud Platform (GCP)
  - gs://**bucket name**/**FASTQ**/**project name**/**synapse ID**/
  - gs://**fc-secure-0a879173-62d3-4c3a-8fc3-e35ee4248901**/**FASTQ**/**igvf\_b01\_LeftCortex**/**syn51669616**/

## Under DASHBOARD

### DASHBOARD

#### CLOUD INFORMATION

Cloud Name	Google Cloud
Location	us-central1 (Iowa)
Google Project ID	terra-06940db2
Bucket Name	fc-secure-0a879173...
Estimated Storage Cost Updated on 6/27/2023	\$0.44
Bucket Size Updated on 6/27/2023	22.2 GiB
Open bucket in browser ↗	
Open project in Google Cloud Console ↗	

## Under DATA / FILES

WORKSPACES

DASHBOARD DATA ANALYSES WORKFLOWS

IMPORT DATA

FILES

TABLES

Search all tables

Team\_0 (1)

Variant\_Zamboni... (9)

REFERENCE DATA

No references have been added.  
Add reference data

OTHER DATA

Workspace Data

Files

ENCODER\_FASTQ/

FASTQ/

JSON\_example/

annotations/

biocontrol/

corebooks/

submissions/

whitetail/

# How to run - Step 1

- Make a copy of the script from the Anvil workflow OR from git to your local machine

The screenshot displays the Anvil Workspaces interface. At the top, there's a navigation bar with a menu icon, a DNA helix logo, 'POWERED BY Terra', a 'BETA' badge, and the text 'WORKSPACES'. To the right of this bar, it shows the current workspace path: 'Workspaces > DACC\_ANVIL/IGVF Single-Cell Data Processing Jamboree > Analyses'. Below the navigation bar is a tabbed interface with 'DASHBOARD', 'DATA', 'ANALYSES' (selected), 'WORKFLOWS', and 'JOB HISTORY'. The main content area is titled 'Your Analyses' and includes a '+ START' button. It contains a table with two columns: 'Application' and 'Name'. The table lists two Jupyter notebooks: 'download\_data.ipynb' and 'tanjin\_download.ipynb'. A context menu is open for the first notebook, showing options: 'Open preview', 'Edit', 'Playground', 'Make a copy' (highlighted with a yellow circle and an arrow), 'Copy to another workspace', 'Copy analysis URL to clipboard', 'Rename', and 'Delete'.

Workspaces > DACC\_ANVIL/IGVF Single-Cell Data Processing Jamboree > Analyses

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

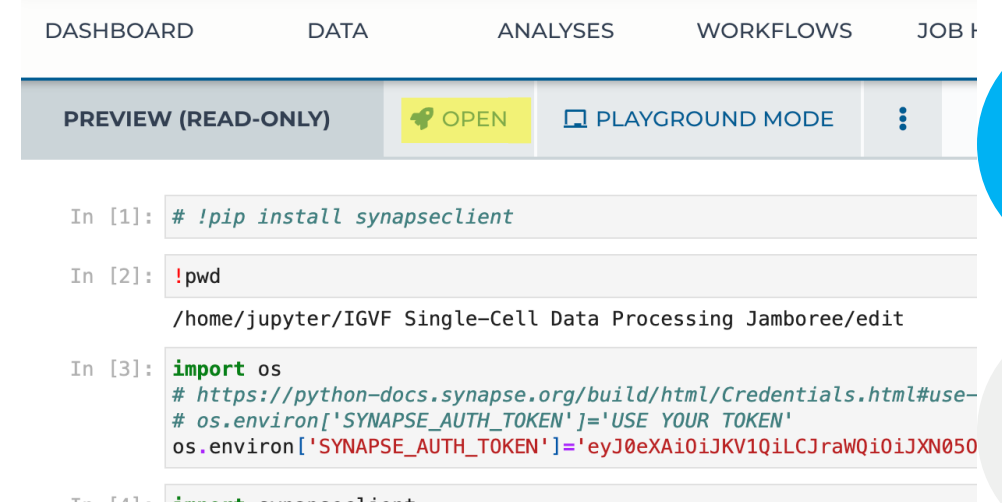
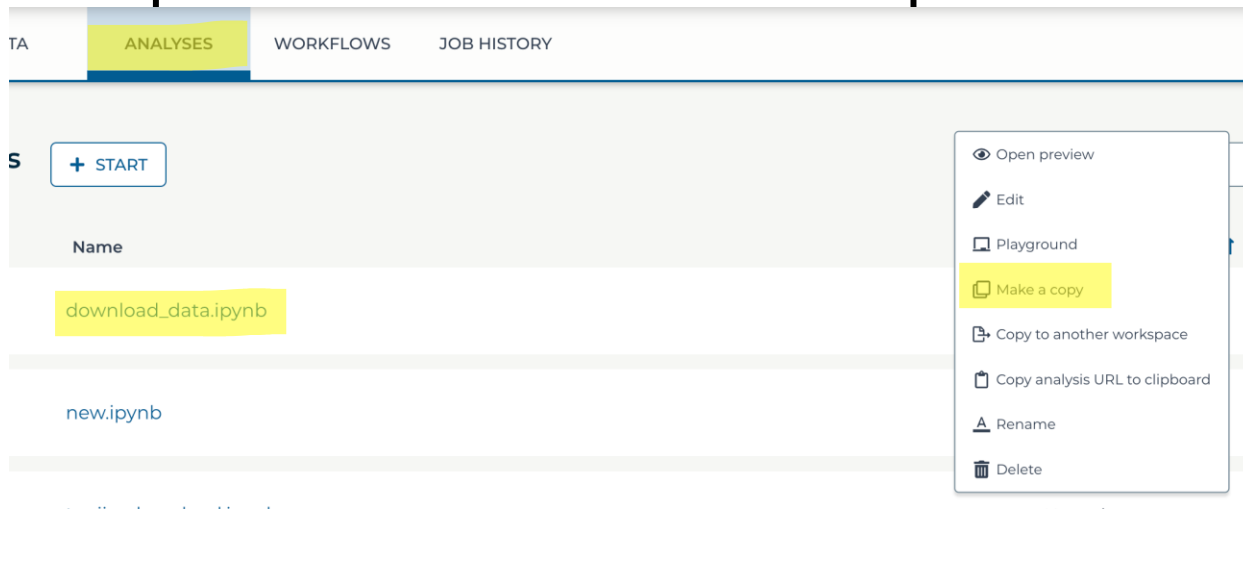
Your Analyses + START

Application	Name
Jupyter	download_data.ipynb
Jupyter	tanjin_download.ipynb

- Open preview
- Edit
- Playground
- Make a copy
- Copy to another workspace
- Copy analysis URL to clipboard
- Rename
- Delete

# How to run – Step 2 - Make your own copy for the script

- Make a copy for the script and rename it: XXXX\_download\_file.ipynb
- Open the file and click on open



# How to run – Step 2

## – Update variables

- Search for TODO on the script

Synapse ID to copy from

SynID syn51669616   
Access 

```
# TODO: change the dataset id
dataset_ids = list(set(["syn51669616", "syn51730625"]))
print(len(dataset_ids))
```

Create local folder  
and look for the  
following code to  
change

```
# TODO: change this to a local folder that is on your machine
local_file_path="/home/jupyter/eila/"
```

Project name to use  
under FASTQ

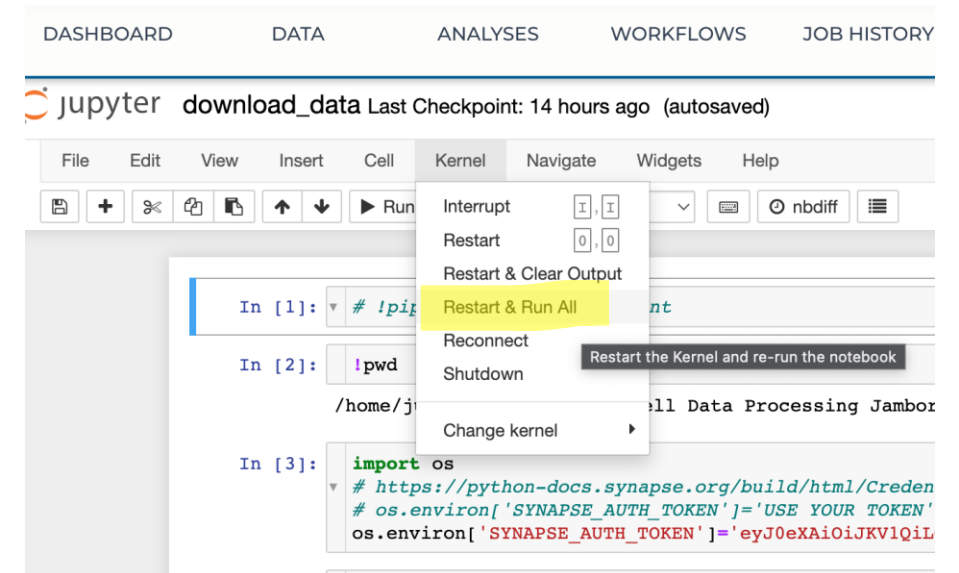
 [igvf\\_b01\\_LeftCortex/](#)

```
# TODO: change here the name of the dataset that you are downloading. For example: "igvf_b01_LeftCortex"
gs_file_path = os.path.join(gs_root_file, "igvf_b01_LeftCortex", dataset_id, dst_file_name)
```



# Execute & Compare

- Compare the files
  - <https://www.synapse.org/#!/Synapse:syn51669616> (synapse – source)
  - [https://anvil.terra.bio/#workspaces/DACC\\_ANVIL/IGVF%20Single-Cell%20Data%20Processing%20Jamboree/data](https://anvil.terra.bio/#workspaces/DACC_ANVIL/IGVF%20Single-Cell%20Data%20Processing%20Jamboree/data) (under workspace data - destination)
- Sometime the Jupyter is not stable and will not complete the upload
  - Just rerun it!
  - The code will not repeat upload that was already completed



**Let me know if you have any  
issue**

