centre for population genomics

murdoch children's research institute

Garvan Institute of Medical Research

Uploading your data to CPG's cloud

A quick guide

Table of Contents

1. Background

i. Purpose

The purpose of this document is to provide easy to follow instructions on how to transfer your genomic data to the CPG's cloud storage.

ii. The Google Cloud Platform

The Centre for Population Genomics currently uses Google's cloud infrastructure to securely store data before it is ingested into Seqr and our metadata platform.

Each project is compartmentalised into its own corner of the Google cloud platform and named accordingly. Projects are split up into "buckets", which is Google's term for storage disk.

In this document, outlines how you can upload your data into the "upload" bucket of your project.

iii. Requirements

To successfully upload your data to the cloud bucket, you will need to install two services using the command line.

1. The Google cloud service for accessing the cloud: `gcloud`
2. The Google storage utility for accessing the buckets: `gsutil`

- **For large uploads (i.e., large batches of sequence data) you should access gcloud through your organisation's server / HPC where your data is stored.**
- **If gcloud is not already installed in your organisation's environment, coordinate with your systems administrator to install and activate gcloud.**

To install gcloud and gsutil, follow these steps.

➢ If you are using Mac OS or Linux, you will need to open a terminal session and copy and paste the commands from the instructions in the above link.
➢ If you are using Windows, you can download an executable installer and follow the prompts to install.

Once you have installed gcloud, you will need to run the command `gcloud init` in a terminal window to successfully start the Google Cloud services command line interface.

2. Authentication

You will have been provided with a service account authorization key. This key, a json file shared via google drive gives you the permission to upload your data into the bucket for your project.

DO NOT share your key with any unauthorized person.

The key gives the holder full access to any data in your projects upload bucket. Store it in a secure location.

If your key is compromised in any way, please inform the CPG ASAP.

1. Download the key from the Google drive you were provided via email.

   The key should have a name like:
   "`your-project-upload.json`" or "`your-project-shared.json`".

2. Run the below command to activate your service account key.

```
gcloud auth activate-service-account --key-file your-key-name.json
```

   Replace the text "`your-key-name`" with the filename of the key you just downloaded. If the key is stored in a different directory you will need to use the full path to the key, e.g.: `home/to/key/your-key-name.json`

3. Successfully activating your service account key will produce the output:

```
Activated service account credentials for:
[main-upload@your-project.iam.gserviceaccount.com]
```

4. Upload your data

To upload your data, use the gsutil *copy* command "`cp`":

```
gsutil -m cp -r source destination
```

Or use the gsutil *remote sync* command "`rsync`":

```
gsutil -m rsync -r source destination
```

***For example**:*

```
gsutil -m cp -r /path/to/data gs://cpg-your-project-upload/subdir/date/
```

- Replace `your-project` with the name of your project
- Replace `subdir` with the upload directory specified in the email
- Replace `date` with the upload date (e.g. "`2023-01-01`")

Note: Include the `-m` flag to upload your data faster by using parallel processing.
Include the `-r` flag to recursively upload all directories in your data folder.

5. Finishing your transfer

A successful upload should result in the output:

```
Operation completed over n objects/xyz B
```

Note: `n` is the number of files in your uploaded folder
`xyz` is the total size of all the files uploaded

6. Getting help

If you require assistance with the above steps, contact CPG's data ingestion coordinator Edward Formaini: edward.formaini@populationgenomics.org.au