HUMAN PANGENOME REFERENCE CONSORTIUM

Data Submission

Introduction

This document provides instructions for submitting data to the Human Pangenome Project (HPP). It describes the steps to structure your data for upload, install the SSDS tool, execute the SSDS command to upload your submission, and how to notify the HPP Data Wrangler of your completed submission.

Ensure That Your Data Is Structured Properly

The HPP accepts data as submissions (batches of data for upload). Please format the directory that you will upload at the sample level (include all data for a given sample under a directory with the sample name from 1000G/HapMap). For large files, please include an md5 checksum as a file with *.md5* appended to the original file name.

Install SSDS

The HPP uses a Python tool, SSDS, to upload data as submissions to the HPP S3 submissions bucket. SSDS can be installed with pip (or pip3) with the command:

```
pip install git+https://github.com/DataBiosphere/ssds
```

More information about SSDS can be found at https://github.com/DataBiosphere/ssds

Store Your Access Keys

In order for SSDS to be able to upload your data to S3, it has to be able to find your S3 credentials. If you have not already been given S3 credentials, please email juklucas@ucsc.edu to request them. Your access keys must be stored in ~/.aws/credentials in the form:

```
[default]
aws_access_key_id=XXXXXXXXXXXXXXXXXXXX
aws_secret_access_key=YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY
```

Note: If you already have aws credentials, you may store your credentials under a named profile. If you are storing your HPP credentials under a profile that is not default, be sure to enter the following command before uploading:

```
# Only necessary if you use a non-default profile name
Export AWS_PROFILE=my_profile_name
```

Upload Your Submission

You can upload your data using the following command:

```
ssds staging upload \
    --submission-id my_submission_id \
    --name my_cool_submission_name \
    /local/path/to/my/submission
```

Where:

- *--submission_id is a universally unique identifier (UUID) provided by the Data Wranglers at the HPRC. If you do not have a UUID, email juklucas@ucsc.edu to obtain one.*
- *--name is a human readable name to make your submission easy to identify. It is recommended to include your institution name and the datatype you are uploading (for example: "UCSC-HiC").*

SSDS will print the files it is uploading to stdout. You can also view the uploaded files in:
https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/

If you would like to add or replace files, you can omit the *--name* argument to update an existing submission. Note that in this case all data in your local directory will be uploaded. Preexisting data in S3 that is in upload directory will be overwritten file by file.

Notify the Data Wrangler of Your Submission

After the submission is complete, please notify juklucas@ucsc.edu.

To assist in automating data upload QC, please include the following information in your email:

- *List of samples uploaded*
- *Sample to library/processing ID mapping (if necessary)*
- *List of file types per sample (the output of a ls command for one sample is sufficient)*
- *Number of files of a given type per sample (if number is not 1)*