Introduction

As part of the Allen Ancient Genome Diversity and John Templeton Ancient DNA Atlas Projects, we are making available to the community a pre-publication set of high-quality shotgun sequencing data from 216 ancient individuals.

While the data for 212 of these 216 individuals are not previously published, the shotgun sequencing was in every case performed on ancient DNA data libraries for which there has been previously published in-solution enrichment data on 1.24 million SNPs. Many of these libraries were generated as collaboration with other ancient DNA laboratories including Ron Pinhasi (University of Vienna, Austria), Wolfgang Haak and Johannes Krause (Max Planck Institute, Germany), Lars Fehren-Schmitz (University of California, Santa Cruz, USA), and Bastien Llamas (University of Adelaide, Australia).

We thank the Paul G. Allen Foundation and the John Templeton Foundation for the funds to generate this unique resource.

Usage

We are releasing the raw data including the bams as a resource to the community. However, please observe the Fort Lauderdale principles, which entitle the data producers to make the first presentation and publish the first genome-wide analysis of the data. The data can be used freely for studies of individual genes or other individual features of the genome, prior to any publication of a manuscript which we hope to have ready within the next year.

If you plan to use some of these data in a publication, please cite the paper(s) in which the sequenced libraries were originally published (as specified in the metadata (https://reichdata.hms.harvard.edu/pub/datasets/agdp/AGDP.metadata.xlsx), along with this website as a reference for the shotgun sequence data (https://reich.hms.harvard.edu/ancient-genome-diversity-project).

Please write to Swapan 'Shop' Mallick [Swapan_Mallick@hms.harvard.edu], David Reich [reich@genetics.med.harvard.edu], and Adam Micco [Adam_Micco@hms.harvard.edu] with any suggestions for improvements. This is only a first release. In future releases, we plan to release more data, and integrate published data from other ancient DNA groups.

Access

We are using the data transfer service, Globus, as a means to distribute the raw data associated with this project. In order to download files, you will need a Globus ID. If you do not already have one, you will need to go to www.globusid.org, click the “create a Globus ID” link, and follow the ID creation process.

Once you have created your Globus ID, please provide it to Michelle Lee [Michelle_Lee@hms.harvard.edu], Adam Micco [Adam_Micco@hms.harvard.edu], and Swapan ‘Shop’ Mallick [Swapan_Mallick@hms.harvard.edu] in order to be granted access to the ReichLabPublic endpoint. You will receive a notification email sent to the address you provided during the Globus ID setup process when you have been granted access.

Using Globus

Below is a detailed description of the steps necessary to setup Globus, gain familiarity with the interface, and download data. While the steps we outline are quite verbose, we anticipate Globus to be relatively straightforward to use after initial setup.

Once you've been granted access to the endpoint, Globus makes accessing and transferring the shared data very simple. Simply navigate to www.globus.org and sign in using your Globus ID and password. You will be directed to the File Manager where you can view the directory structure and initiate transfers from the ReichLabPublic endpoint to your local machine or directly to institution's cluster.

NOTE: The ability to transfer data directly from the ReichLabPublic endpoint to your institution's cluster without first downloading to a local machine is dependent on your institution having setup a Globus endpoint. If you are unsure if your institution has a Globus endpoint or if you have questions about finding it, please contact your research computing/high-performance computing representative for help.

Finding the ReichLabPublic endpoint

From the File Manager, click on the Collections search box on the upper left of the page. You will be directed to the Collection Search interface where you should navigate to the "Shared with You" tab. From here, locate and select the ReichLabPublic endpoint. Once you've selected this, you'll be redirected back to the File Manager where you will see the top level directory of dataset's directory structure on the left-hand side.

Setting up a personal endpoint

In order to transfer data to your local machine from the Reich Lab endpoint you will need to set up a personal endpoint using the Globus Connect Personal client. This client is available for Windows, MacOS, and Linux and enables transfers of data from Globus endpoints to a location on your local computer.

From anywhere in Globus, select the Endpoints tab in the sidebar and click the "Create a personal endpoint" link in the upper right corner of the page. You will be directed to a download page for your local operating system. Download and run the installer following the onscreen instructions. Once the you complete the install process, you will be prompted to log in. Clicking the Log In link will open a browser window where you will need to sign in using your Globus ID and authorize Globus traffic to/from your local machine. Click Allow on this page to proceed.

Once you've allowed Globus traffic, you will be presented with a form asking for the following information.

- **Owner Identity**: Do not change this — it will default to your Globus ID when you log in.
- **Collection Name:** This will be the name of your personal endpoint and will be what you use as a target for data transfers. We recommend using a descriptive name such as "work-laptop."
- **Description:** This field is optional — use it to differentiate endpoints if you would like.

Clicking “Save” will confirm these details and launch the Globus application in the background. Your new personal endpoint will now be available for use on https://app.globus.org/endpoints.

Initiating a data transfer

From the Globus File Manager, specify the ReichLabPublic endpoint in the left-hand Collection field and your personal or institutional endpoint in the right-hand Collection field. You should see the top level directory listing available ReichLabPublic endpoint projects on the left and your local home directory on the right.

Navigate to the directory containing the files you would like the download using the left side of the File Manager and to your destination directory using the right side. You can also specify these locations using the Path fields. Once you’re in the correct directories, select the files and/or directories you would like to transfer and initiate the transfer by clicking the Start button with the right facing arrow. You will see a pop-up confirming the transfer has started in the top right of your screen. You will also receive an email when your transfer completes or if it fails for any reason.

To monitor the status of your transfer, select the Activity tab from the Globus sidebar. Here you will see a list of all current and recent transfers. Select any of these to view the status of any in progress transfers or the details and log of past ones.

ReichLabPublic Endpoint Directory Structure

The ReichLabPublic endpoint follows the following organizational hierarchy:

- **Project** (e.g. agdp_release)
  - **Version** (e.g. v3.5b)
    - **Data File Format** (e.g. bams)
      - **Sample** (e.g. l2978)
        - **Data files**

Details on each of these levels can be found below.

PROJECT

Opening the ReichLabPublic endpoint in the Globus File Manager will take you to the top level directory containing subdirectories for each of the released public datasets. At this time (February 2021), agdp_release will be the only subdirectory present. However, it is our hope to use Globus as a mechanism to improve data accessibility for collaborators and the public at large so we expect to add to this list of available datasets in the future.

VERSION

Datasets in the ReichLabPublic endpoint follow the below versioning schema:

v[*major version #*].[*minor version #*][*revision letter*] (e.g. v3.5b)

As datasets grow and samples are updated, we will periodically release new versions of released datasets. Depending on the changes made, we will bin these changes into either a

major version, a minor version, or a revision. Examples of changes that may be included in each are below:

- **Major Version -** Samples added/removed
  - **Minor Version -** Additional coverage added on existing samples
    - **Revision -** Bug fixes

Every minor version and revision within the current major version will be available directly within the project directory. With each major version release, we plan to retain all previous releases but move them to the archive subdirectory.

DATA FILE FORMAT

Within each version, we then group released data by file type to make downloading homogenous datasets for analysis straightforward.

SAMPLE

Within a file format grouping, we then group the data files both by sample in subdirectories named for their Reich Lab IDs and all together in the all subdirectory.

Questions and Concerns

While we have high hopes for the Globus platform to become an avenue for rapid data sharing and collaboration, we are new to it and are sure to hit some bumps along the way. If you encounter any issues using the ReichLabPublic endpoint or spot any problems with the shared datasets, please write to Adam Micco [Adam_Micco@hms.harvard.edu] and Swapan 'Shop' Mallick [Swapan_Mallick@hms.harvard.edu].

We are aware that there are many types of analyses data such as these might be used for, and it is hard to anticipate some of these. We welcome any observations, and/or suggestions for simplifying/improving the download process.