H3ABioNet Data Submission Guidelines

H3ABioNet Data Submission Guidelines

This document describes the submission of H3Africa project data to the H3ABioNet Data Archive.

1. General Information

The H3ABioNet data archive team (HDT) has been tasked with assisting the H3Africa data generating projects to submit their genomic and phenotype data to both the H3Africa data archive and the European Genome Phenome Archive or other relevant public repositories (such as the European Nucleotide Archive) in compliance with the H3Africa Data Sharing and Release Policy. H3ABioNet will function as a data-coordinating center for the submission of H3Africa genomic data and conduct basic validation checks to ensure that the data submitted is a publishable dataset and is what was initially intended to be sent by agreement with the data submitter. Some of the information in these guidelines have been extracted from the EGA documentation (https://ega-archive.org/).

2. Overview of the H3Africa data submission process

H3Africa projects that are almost ready to submit their data should contact the H3ABioNet Data Archive team via the following email address: archive@h3abionet.org to signal their intention to submit at least six weeks prior to the actual submission.

The HDT will contact the data submission requester to arrange a meeting which will encompass completion of a sample Data Submission Request (DSR) form, the modalities of the data submission that includes the types of data, the estimated data size, number of samples, encryption and methods of data transfer. The completed DSR will be provided to the data submission requester to confirm the details. Once all parties have agreed, the submission will be logged onto the H3Africa Data Submission dashboard and a data submission pack will be provided to the data submission requester.

3. Types of data to be submitted

All files must be encrypted before submission. The H3Africa data archive will accept data closely aligned to EGA specifications and includes phenotype, sequence data (human and microbiome) and array data as summarized below. Non-human sequence data can be submitted to the European Nucleotide Archive.

Files that will be requested upon submission

The exact list of file types that will be requested depends on the type of study submitted. However, all H3Africa data submissions will have the following files in common when submitting their data:

- A summary description of the study in text format. This should include the study type e.g case / control, cohort, microbiome and characteristics under investigation.
- A summary description of the dataset in text format. This should include the types of data (NGS, chip, microbiome), platform used for generating the data and the number of samples.

H3ABioNet
Pan African Bioinformatics Network for H3Africa

H3ABioNet Data Submission Guidelines

- A blank specimen of the consent form used for collecting the data. **Please do not submit data for participants who did not sign the consent form or within the form did not consent to data sharing.**
- A blank specimen of the Case Report Form used to collect the data.
- An Institutional ethics code and the name of the Institute that provided ethics approval for the study.
- A mapping file that indicates the structure of the data submitted.
- A Phenotype file that indicates the Phenotypes of the data submitted.

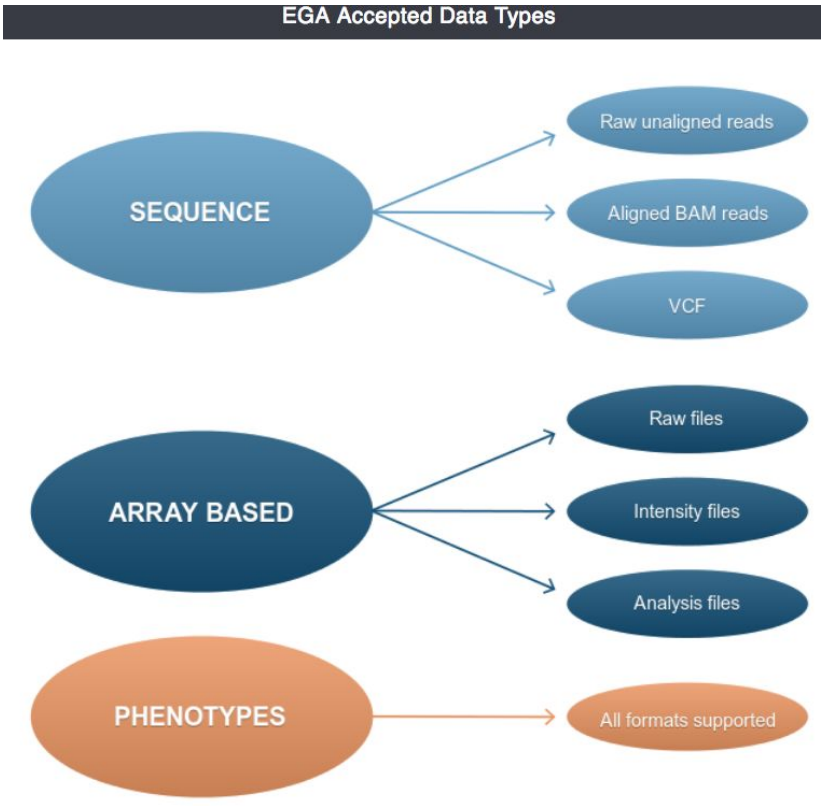

Figure 1. Data types accepted by the EGA

4. Sequence data (exome, WGS, microbiome)

The following files should be submitted for NGS sequence data pertaining to exome, whole genome sequence or microbiome data.

4.1. Fastq sequence data:

H3ABioNet Data Submission Guidelines

Fastq files should be linked with de-identified participant IDs and should be de-multiplexed prior to submission so that each experiment is submitted with files containing data for a single sample only.

Single and paired end reads as Fastq files that meet the requirements below will be accepted.

- Quality scores must be in Phred scale. For example, quality scores from early Solexa pipelines must be converted to use this scale. Both ASCII and space delimited decimal encoding of quality scores are supported. We will automatically detect the Phred quality offset of either 33 or 64.
- No technical reads (adapters, linkers, barcodes) are allowed.
- Single reads must be submitted using a single Fastq file and can be submitted with or without read names.
- Paired reads must be split and submitted using either one or two Fastq files. The read names must have a suffix identifying the first and second read from the pair, for example '/1' and '/2' (regular expression for the reads "^(.*)([\\.|:|/|_])([12])$").
- The first line for each read must start with '@'.
- The base calls and quality scores must be separated by a line starting with '+'.
- The Fastq files must be compressed using gzip or bzip2.

4.2. BAM files:

All BAM files should be linked with de-identified participant IDs and must be de-multiplexed before submission although multiple BAMs as part of an analysis can be submitted. All BAM files must be readable with SAMtools and Picard. Colour spaced BAM files will not be accepted.

4.3. CRAM files:

All data files should be linked with de-identified participant IDs and must be de-multiplexed before submission so that each experiment is submitted with files containing data for a single sample only. All CRAM files must be readable with SAMtools and CRAMToolkit and the reference sequences must be present in the CRAM Reference Registry. The ArchiveCRAM specification outlines the requirements for BAM and CRAM submissions.

4.4. VCF files:

VCF files should be submitted where applicable. Please ensure validation of the VCF files for submission by using the EVA VCF validator or minimally the VCF Tools validator.

4.5. Microbiome data:

In addition to the above mentioned Fastq file specification listed, the final analysis BIOME file or at minimum, the OTUs should be submitted as separate files within the submission. The sequencing platform for the microbiome data should be included in the data set description upon submission.

5. Genotyping array data

H3ABioNet
Pan African Bioinformatics Network for H3Africa

H3ABioNet Data Submission Guidelines

The raw intensity files (e.g .CEL for Affymetrix or iDATs for Illumina) linked with de-identified participant IDs. The chip model and platform name should be included within the dataset description. A manifest file that describes the SNP or probe content on the array, and the name of the software (including version) used for calling the genotypes should also be submitted. The final reports and analyses files generated for the study should be included.

6. Mapping files

These are files that detail how all the files intended for submission relate to each other in terms of the project. The HDT member will work with the submitter to ensure that these files contain the necessary information to meet H3A and EGA archive requirements.

7. Phenotype data

The data submitter should provide as much phenotype data as possible to make the data useful to other researchers. Phenotypes can be Gender, Ethnicity, Country, and any phenotype measurements collected within the Case Report Forms (CRFs).

8. Validation

Validation is the process of ensuring that the data submitted is what the data owner / submitter agreed to send to the archive. The below validation checks are run:

- Do the checksums match?
- Are there mapping files present?
- Do the number of samples match what is expected?
- Are all the files present for each de-identified participant ID?
- Is there a mismatch between participant IDs and files?
- Do all the files have phenotypic data present? what field names are missing? what should be collected?
- Is there a dataset summary description and study abstract present?

9. Data submission pack

The submission pack will contain the following:

- A pre-populated DSR for the user to confirm and fill in the missing information regarding the submission details.
- Encryption instructions.
- Globus Installation instructions.
- Phenotype File Template
- Mapping File Template

10. Data transfer methods

H3ABioNet Data Submission Guidelines

H3ABioNet will employ two types of data transfer methods. The primary data transfer method will be via the Globus Online (GO) application. GO is designed to securely and reliably transfer large data files across the internet. All data should be encrypted for transfer.

10.1. Online Transfer Method

- The data submitter should notify the HDT of an impending data submission at least six weeks in advance.
- Data will be submitted by individuals via a secure data transfer mechanism agreed upon between the H3ABioNet data team and the submitter. The current recommended application for electronic data transfer is Globus Online.
- The data submission logs will be recorded and attached to the high-level folder containing the complete data submission. The transfer logs will be sent to the submitter as confirmation.

10.2. Courier Workflow Method

- For areas where the Internet is too slow or unreliable for electronic transfer, the HDT will ship portable hard disks to and from the submitter via registered courier.
- The data submitter will be required to furnish HDT with the relevant contact details pertaining to the courier delivery allowing the HDT to track the hard drive while in transit- in the event the data submitter commissions the courier. In most instances, the HDT will arrange for delivery of the hard drive/s.
- Once the data has been transferred to the drive/s, the HDT will arrange for the collection of the drives.
- Due to the limited number of hard drives, the individual should move their encrypted data to the physical hard drive/s and work with the HDT to arrange timely collection of the hard drive for courier to the CBIO offices in Cape Town. Failure to do this would result in the individual or data owner incurring the cost of the hardware.

11. Data Security

To minimise risk to the submitted data, the following measures are employed.

11.1. Encryption

All data submitted to the H3Africa Archive needs to be encrypted.

The public key will be given to the data submission requester as part of the submission pack. They will follow instructions provided in the submission pack on how to encrypt the data using an encryption tool called GPG. The H3ABioNet Data Archive team will only accept the data if it is encrypted, otherwise we can't be responsible for its security.

The private key is stored securely by the H3ABioNet Data Archive team.

11.2. Security Best Practices

- The physical hardware housing this data will be stored in the UCT Data Centre (UCT DC)
- The data will remain in the encrypted state all the time, except when undergoing validation in a dedicated secure environment. Access to this computing environment is restricted to only the HDT members that will validate the data. After validation is complete, the dataset is encrypted again before moving to long-term storage.
- The hardware housing the archived data will reside in the access controlled UCT DC which is protected by swipe card access in a lockable network cabinet.
- Physical access to this room is limited to core IT and maintenance personnel only.

11.3. Disaster Recovery

The UCT DC employs the following disaster recovery (DR) infrastructure.

- The hardware which holds the physical H3Africa data has been designed with internal physical disk redundancy for automatic failover.
- Each hardware chassis in the archive solution has a minimum of two power supplies connected to separate power outlets.
- Hardware is powered via Inline UPS devices for immediate failover power and an external generator backup power for longer outages.
- Data is replicated and not backed up.
- Dual network cards are installed in all server hardware for load balancing and failover.
- A fire suppression gas infrastructure in the event of fire and a climate control monitoring and notification system is in place.

12. Submission to EGA

The H3ABioNet Data Archive is intended to be temporary archival storage only. The submitted data will in turn be submitted to EGA, which will be the permanent storage. H3Africa and H3ABioNet have agreed on a 9 month holding period between the H3ABioNet Data Archive and the EGA submissions. However, submitted data will remain archived up until 2022 in encrypted form on the H3Africa data archive long-term archival storage platform.

- The HDT team will only accept a submission into the archive when it has passed the validation checks as stipulated in the submission pack.

H3ABioNet Data Submission Guidelines

- The HDT will obtain unique identifiers (accessions) from EGA, in order to submit data to this repository, and store a mapping to the data they identify.
- The HDT will interrogate submitted datasets to ensure that they meet the requirements specified by H3Africa and EGA, and work with submitters to get the data into appropriate format where needed.

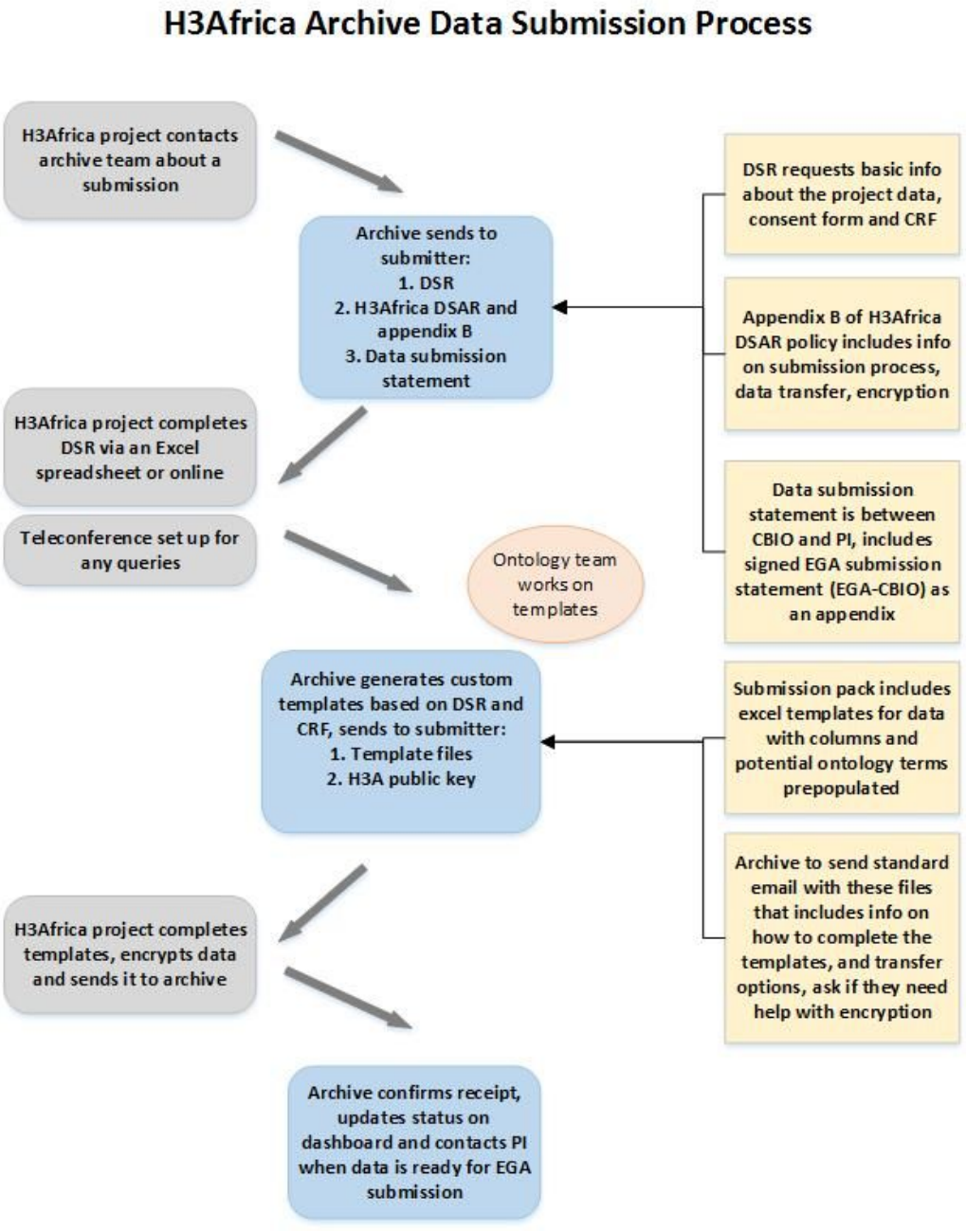

Figure 2. Overview of the data submission process

H3ABioNet
Pan African Bioinformatics Network for H3Africa

H3ABioNet Data Submission Guidelines

13. EGA terminology

XML file name	Description
Submission XML	Describes the submission transaction, contact details, md5 checksum values before and after encryption
Study XML	Describes study in detail, title, study name and abstract and provides unique identifier / accession
Sample XML	Description of each of the samples used in the study
Experimental XML	Experimental details such as library preparation, sequencing platforms and type
D(B)AC XML	Description of the data access policy and url
Policy XML	Describes the data access agreement to be linked to the D(B)AC
Run XML	Describes data file and relation to experiment
Analysis XML	Describes files generated from experiments e.g BAM, VCFs
Dataset XML	Describes the data files that constitute the dataset and linked to the specific policy in place

H3ABioNet Data Submission Guidelines

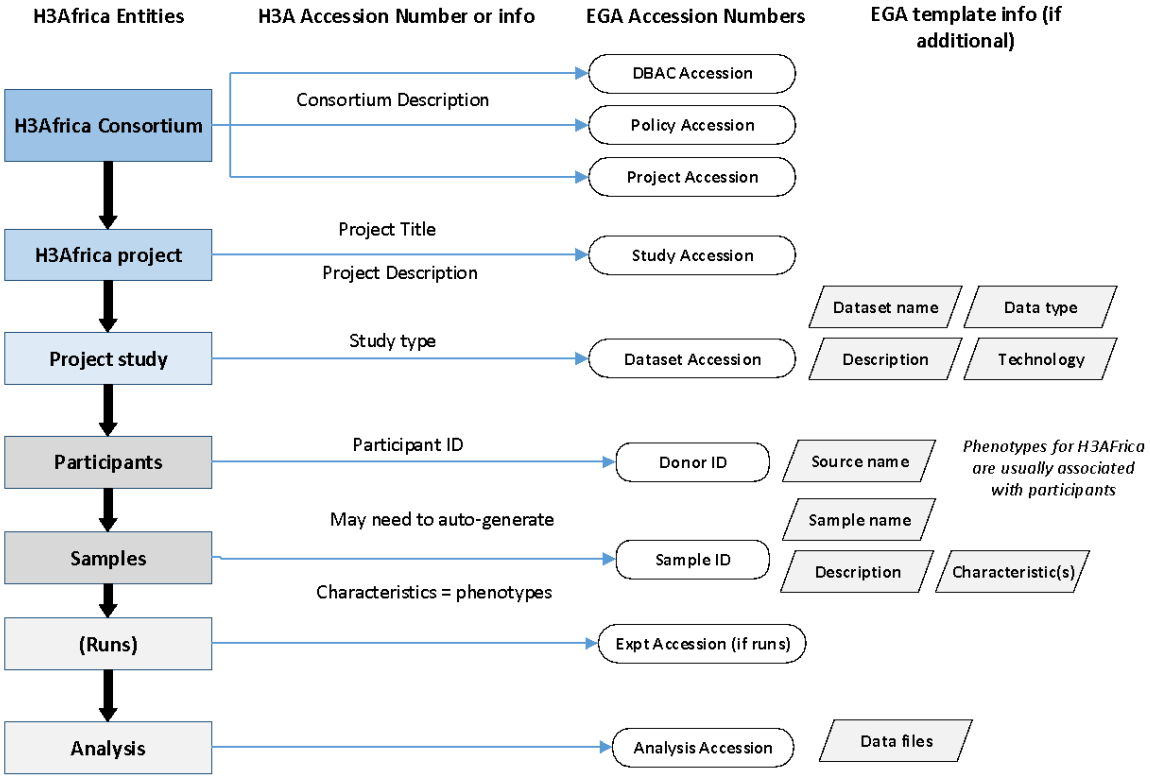

Figure 3. mapping of data entities between H3Africa projects and the EGA

14. H3Africa Catalogue

The H3Africa catalogue is an initiative to make the H3Africa metadata submitted to the H3Africa Archive for EGA and metadata from samples in the biorepositories available via a public interface. The H3Africa catalogue is a public web interface for researchers to search high level phenotypic data that projects have submitted to the H3Africa Archive and EGA. To download the data, researchers will need to request data from EGA via the DBAC.

14.1 Metadata submitted from the Archive to the Catalogue

The following describes which project data, where available, will be made available for searching in the Catalogue. Note, participant and biospecimen identifiers will be hidden from public view in the catalogue. Biospecimen IDs will only be made available to the H3Africa secretariat and DBAC as required for processing of data and biospecimen access requests. The H3ABioNet Data Archive Team will be responsible for submitting selected metadata from H3Africa projects to the Catalogue.

Data that will be made available will be dependent on the data the projects submit. The data that the archive intends to collect and make available for searching in the catalogue (if available) is listed below:

1. Project information (title, abstract, study design)
2. Project consent code
3. Anonymized Participant ID (internal, not visible to public)

4. Specimen id (internal, not visible to public)
5. Age at Baseline (in years)
6. Gender/Sex
7. Ethnic group (self reported)
8. Height (in millimeters)
9. Weight (in kg)
10. BMI
11. Disease phenotype (or control)
12. HIV status if collected (internal, not visible to public)
13. Files type generated by the study

For the following we will only submit information on whether such data is being collected by the project in any form, no values will be submitted for individuals

1. Education
2. Diet
3. Smoking
4. Alcohol
5. Substance Use
6. Blood Pressure / Hypertension
7. HIV information

15. Contacts

Resource	**Description**
archive@h3abionet.org	Email address used for data submission communications and alerts
CBIO node postal address	University of Cape Town Faculty of Health Sciences Computational Biology Group Room N1.05, level 1 Wernher and Beit Building North, Anzio Road, Observatory 7925, Cape Town South Africa