H3ABioNet Data Submission Standard Operating Procedure

A data submission agreement between H3ABioNet and H3Africa Consortium members

Created by:

Data Management Task-force

Prepared to use internally for:

H3Africa Data Archive Team

Document Control

Date	Author	Authorization	Version	Description
6 Sep 2013	Suresh Maslamoney	Data management task-force	1.0	Initial development of the CBIO data access policy for stored H3Africa research data
29 May 2014	Suresh Maslamoney	Infrastructure Working Group (ISWG)	1.0	Policy accepted at the 29 May 2014 ISWG meeting
22 Jan 2015	Suresh Maslamoney		1.2	Reworded and reflects updated archive design
8 Jul 2015	Danny Mugatso	Data Archive Team	2.0	Incorporated new design and processes
4 May 2017	Ziyaad Parker	Data Archive Team	2.1	Added missing and updated information
25 July 2017	Ziyaad Parker	Data Archive Team	3.0	Combining previous documentation, importing it to more editable and collaborative format in Google format. Filling in remaining information. Changed the name from Policy to SOP.
30 July 2018	Ziyaad Parker	Data Archive Team	3.1	Add more information related to the submission of Sequence files

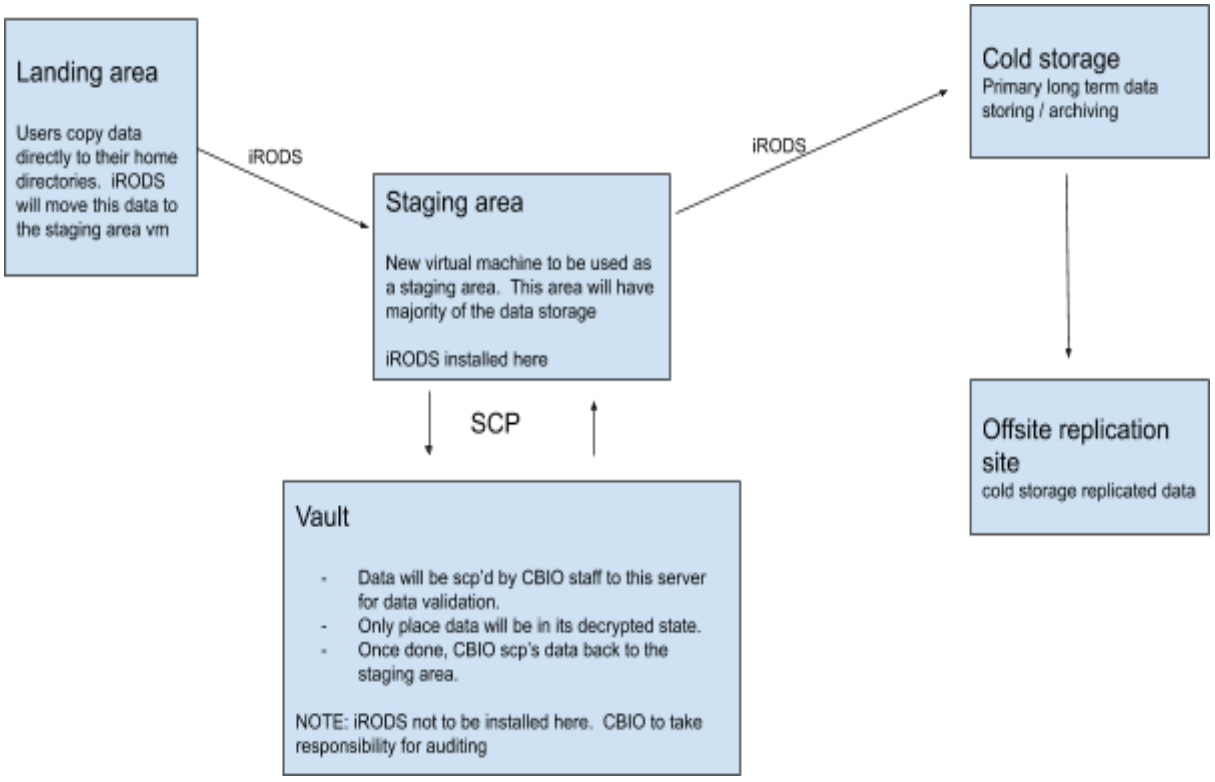
Acronyms

Acronym	**Definition**
DC	Data Centre
DR	Disaster Recovery
DSR	Data Submission Request
CBIO	Computational Biology Group
EGA	European Genome-Phenome Archive
HDT	H3ABioNet Data Team / H3Africa Data Archive Team
OTU	Operational Taxonomic Unit: *"Taxonomic level of sampling selected by the user to be used in a study,…"*
SOP	Standard Operating Procedure
UCT	University of Cape Town
UCT DC	University of Cape Town Data Centre

Table Contents

1. Overview

H3ABioNet has been tasked with securely archiving the genomic and phenotypic research data generated by the H3Africa projects. This data will be stored in a secure data archive solution located at the University of Cape Town, South Africa. In addition to archiving this data, H3ABioNet will, working in conjunction with the relevant H3Africa projects, submit this data to the European Genome-Phenome Archive (EGA), where access can be obtained through the H3Africa Data Access Biospecimen Committee for the project. Data should be submitted to H3ABioNet for archival purposes after it has been through the relevant quality control processes.



2. Data Submission Process

H3ABioNet does not own any of the H3Africa research data submitted to the archive, and as such will only interrogate it during the validation phase to confirm that it meets EGA and H3Africa requirements. If the dataset passes this validation, a description of the dataset including the contact details of the relevant data owner will be made available. The EGA box number is: ega-box-617.

Due to the consistent changes and time required to maintain the internal Archive dashboard. Email and Google Drive will be used in the interim process. The data will be populated onto the internal dashboard when ready.

1. User (PI or Data Submitter) emails archive@h3abionet.org,
2. The H3ABioNet Archive Team (HDT) will respond by setting up a teleconferencing call to facilitate the requirements with the data submitter. The agenda and checklist will be to discuss the DSR and how to submit the data, policies and statements. The HDT team will need to take notes and fill it in.
3. HDT will pre populate the DSR with as much information as possible.
4. A method for transferring the data will be agreed upon. The default transfer mechanism is via the Globus Online (GO) application as described in the Data Transfer Section.

Each submitting H3Africa Project must keep a local copy of their data at all times, even after submission. The H3ABioNet Data Archive is not intended to function as a backup. The data contained within it will not be searchable, accessible or retrievable, except by HDT members for the purposes of formatting and submitting to EGA. To also make it available via the Online Catalogue.

2.1 Data Submission First Contact

At the first teleconference call the HDT will discuss the following details, it will be zipped into one folder:

1. Appendix B (Globus and Encryption - public key)
2. DSR
3. Submission Statements (EGA and H3Africa Submission Statement)
4. H3Africa (DSAR) Policy documentation

The Data Submitter will need to respond with all the information in one email if possible. To go ahead with data transferring process they will need to sign the H3Africa Archive Submission Statement and fill in most of the information pertaining to the DSR. Required are marked in red.

2.2 Data Submission Request

2.2.1 Information

- Project Name
- Project Acronym
- Project PI / Stakeholder
- Abstract/Description for the Project / Study

- Description for the Data Set
- Institutional reference ethics code / number
- Number of Samples
- Number of Cases
- Number of Controls
- Other (e.g: suspect cases, latent infections, etc)
- Platform Type (ie: Illumina, Cardio Metabochip, H3Africa Chip, Affymetrix)
- Type of data for submission (ie: FastQ, BAM, VCF, etc)
- Additional files are in the submission (ie: mapping files, phenotype data, etc)
- How is the data linked, if there are any
- Estimated overall size of the data submission
- Type of Study (ie: GWAS, microbiome, longitudinal, cohort, case-control, trio)
- Date intended to submit the data
- Estimated date to submit the data
- Link to Github for the code used to generate the data
- Any other information not included

2.2.2 Files/Attachments

- Blank copy of the CRF/Questionnaire used to collect the data
- Blank copy of the Consent Form used to collect the data

2.3 Submitted Files Minimum Requirements

The files that are submitted to H3ABioNet should be in line with those required by the EGA (see separate document on EGA submission guidelines at https://www.ebi.ac.uk/ega/). Other data not mentioned in below but mentioned on the EGA will also be accepted.

2.3.1 Genotyping arrays

- Raw intensity files, and any other files necessary for recalling
- Description of workflow (sufficient to reproduce genotype calling)
- PLINK dataset (ped/map or bed/bim/fam) of genotypes used in the analyses

2.3.2 Sequencing projects

- Sequence reads (FASTQ format) or
- Sequence alignments used (BAM format)
- Variants called (VCF file). May be sent later, before EGA submission

2.3.3 Metagenomics

- Sequence reads (FASTQ format)
- OTU tables to be used in publication

2.3.4 Phenotype data

- Any phenotype values collected for the publishable dataset

Participant_id	Gender	Country	Ethnicity	Case-Control	Disease-State
001	Male	South Africa	Asian	Case	Malaria
002	Female	Ethiopia	Black	Control	Malaria
003	Male	Zimbabwe	White	Case	Malaria

2.3.5 Ethics approval documentation

- Stipulations or restrictions imposed by ethics review committees on secondary use of data
- Ethics approval reference numbers, as used in the study to be published
- Relevant documentation on any externally sourced data sets

2.4 Submission Pack

The Submission Pack will contain the following detail:

1. DSR for user to confirm and fill in the remaining information regarding the submission details. The current details being asked for DSR are in Section 2.2
2. See 2.3.1 Genotyping arrays
3. See 2.3.2 Sequencing files

2.3.1 Genotyping arrays

- EF template for the data submitter to fill in the details.

2.3.2 Sequencing files

BAM files

Participant_id	File_name	Date_of_analysis
001	1_001.bam	2016-10-28T00:00:00Z
002	2_002.bam	2016-10-28T00:00:00Z
003	3_003.bam	2016-10-28T00:00:00Z

VCF files

Participant_id	File_name	Date_of_analysis
001	1_001.vcf	2016-10-28T00:00:00Z
001	1_002.vcf	2016-10-28T00:00:00Z
001	1_022.vcf	2016-10-28T00:00:00Z
002	2_001.vcf	2016-10-28T00:00:00Z
002	2_002.vcf	2016-10-28T00:00:00Z

Experiment file

Participant_id	name	strategy	source	selection	instrument_model	construction_protocol
001	H3A_1	WGS	Genomic	Hybrid Selection	Illumina HiSeq 2000	Illumina Truseq PCR-free kit
002	H3A_2	WGS	Genomic	Hybrid Selection	Illumina HiSeq 2000	Illumina Truseq PCR-free kit
003	H3A_3	WGS	Genomic	Hybrid Selection	Illumina HiSeq 2000	Illumina Truseq PCR-free kit

https://www.ebi.ac.uk/ega/sites/ebi.ac.uk.ega/files/documents/Experiment_illumina_paired.xml

Run files

Participant_id	File_name	Date_of_analysis
001	1_001_r1.fastq	2016-10-28T00:00:00Z
001	1_001_r2.fastq	2016-10-28T00:00:00Z
002	2_002_r1.fastq	2016-10-28T00:00:00Z
002	2_002_r2.fastq	2016-10-28T00:00:00Z

2.5 EGA Webin Uploader

3. Data Transfer

H3ABioNet will employ two types of data transfer methods. The primary data transfer method will be via the Globus Online (GO) application. GO is configured to track the data transfer and in the event of a power outage or a break in internet access, GO will continue the data transfer once power or internet access is restored from the point it failed.

3.1 Data Transfer Methods

3.1.1 Online Workflow 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.

For areas where the Internet is too slow or unreliable and electronic transfer is not feasible, an alternative method such as shipping portable hard disks via courier will be arranged.

3.1.2 Courier Workflow Method

- When submitting larger datasets in an unreliable internet infrastructure, it is recommended to use a courier to transfer the data.
- The data submitter should notify the HDT of the impending data submission six weeks in advance.
- HDT will commission a courier company to deliver the physical hard drives to the individual.
- The individual would need to move their encrypted data to this physical hard drive/s and the hardware to the CBIO offices in Cape Town in a timely manner. Failure to do this would result in the individual or data owner incurring the cost of the hardware.
- The data submitter will be required to furnish HDT with all the relevant data pertaining to this delivery allowing them to track the hard disk should the data owner use a courier of their choice. This information would include: Courier company name and contact details, Tracking number, date package sent and place of departure.
- The CBIO postal address is provided in Appendix I

4. Data Storage

The physical hardware housing this data will be stored in the UCT Data Centre (UCT DC) and will implement the following:

4.1 Security Best Practices

- 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.

4.2 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 to a separate DC located at another university within Southern Africa
- 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.

5. Data Flow Processes

The data arrives in the Landing Area in the encrypted format using the instructions provided in in the Submission Pack. The validation gets done inside the Vault. Thereafter it gets moved to Cold Storage.

5.1 Encryption

The H3Africa Archive private key is stored on the vault for submission to the H3Africa Archive. For data submission to EGA, EgaCryptor will be used. See more on this link: https://ega-archive.org/submission/tools/egacryptor

5.1.1 Steps to encrypt/decrypt the data for the African Archive

5.1.1.1 Encryption

The public key will be sent to the data submitter as part of the submission pack. They will follow instructions provided in the submission pack on how to encrypt the data. Document Title: Encryption and copying of the data

5.1.1.2 Decryption

Command: gpg --decrypt test-file.asc

More information on this link: http://linux.101hacks.com/unix/gpg-command-examples/

5.1.2 Submitting data to EGA

For Genotyping Array Data the data submitter will to fill details on the EF Template. HDT will be validate the EF Template. When this has been checked, confirmed and completed it will emailed to ega-helpdesk@ebi.ac.uk,

The software client Aspera will be used to send the submission data to EGA. Depending on the size and speed of the data set the preferably option is to send it from the Landing Area or from personal computers if there arises any delays.

5.1.3 Steps to encrypt the data for submission to EGA

Use EGACryptor (software built by EGA and contains their public key), encrypt the files separately using one command. This can all be done in one folder.

Command: java -jar ../EgaCryptor.jar -file *

More information on this link:
https://ega-archive.org/submission/tools/egacryptor#EncryptMultipleFile

5.2 Globus

5.2 Globus Endpoints

- Heinedej - Landing Area for default receiving of data
- H3data - For Baylor and other custom data submissions

5.3 Validation

5.3.1 Definition of validation?

Validation gets done in the Vault only. Validation is the process of checking data to see what the data submitters say they are going to submit is actually going to be sent using the preferred data transfer methods. Doing checks on the files and making sure it is in line with EGA standards.

Updates for validation will be done on online Google Sheets in each Project Folder. If the PI's want the progress these folders will be shared externally and separately with the PI's.

5.3.2.Validation steps to follow

Move data to the vault, decrypt and validate:

- 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? the field names what is missing? what should be collected?
- Is there a dataset summary description and study abstract present

GWAS specific:

- Is there a plink dataset?
- Does it load?
- Check that sample counts etc match declared counts ids.
- Check are there raw files (cel for Affy, idat for Illumina) for each sample declared
- Is there a phenotype file that plink can read? (optional)

Passes validation:

- Map data to EGA XML/Json schemas
- Encrypt with EGA key (EGACryptor)
- Re-encrypt the data with second H3A key and move to cold storage
- Check data that is missing from CRF or Questionnaire

5. Data Access

The H3ABioNet archive was not intended to be a repository for sharing datasets. As such, after a submission is received, it will not be accessible to anyone other than the H3ABioNet HDT, and this access is solely for submission to the EGA. Groups who require access to a dataset will have to refer to the respective data owner, or apply for access to it via the EGA once available.

6. Data Submission to EGA

The H3ABioNet Data Archive is intended to be a 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 for up to 5 years after submission to EGA.

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

7. Roles and Responsibilities

- H3ABioNet will in no way own the H3Africa archived data. H3ABioNet acts only as a conduit to securely archive the H3Africa data.
- HDT will check that submitted datasets conform to EGA and H3Africa requirements, and make this status available to the submitter, consortium and funders.
- H3ABioNet will not be responsible for ensuring the validity of the data. H3ABioNet will not grant access to this data to any individual or organization other than the EGA.
- The H3ABioNet Data Archive Team are responsible for ensuring that the submitted data is in no way leaked to parties for whom it is not intended.

8. Appendix 1 - Resources

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