H3ABioNet Data Access Policy for H3Africa Data

# Document Control

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Reviewer** | **Version** | **Description** |
| 6 September 2013 | Suresh Maslamoney | 1.0 | Initial development of the CBIO data access policy for stored H3Africa research data |
| 28 September 2013 | Ayton Meintjes | 1.1 |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Acronyms

|  |  |
| --- | --- |
| **Acronym** | **Definition** |
| HDT | H3ABioNet Data Team |
| DAR | Data Access Request |
| DAC | Data Access Committee |
| CBIO | Computational Biology |
| DR | Disaster Recovery |
| DSR | Data Submission Requirement |
| SAN | Storage Area Network |
| EGA | European Genome-Phenome Archive |

Table of Contents

Document Control 1

Acronyms 1

Overview 3

Purpose 3

1. Data Submission Policy 3

1.1 Submitted files minimum requirements 4

2. Data Access Policy for H3Africa Consortium Members 6

3. Data Storage 7

3.1. Security Best Practices 7

3.2. Disaster Recovery 7

4. Data Transfer 8

4.1. Online Workflow Method 8

4.2. Courier Workflow Method 8

5. Responsibilities 9

6. Resources 9

7. Appendix 9

7.1. Data Submission Workflow 9

7.2. H3ABioNet Data Submission Form 11

7.3. H3ABioNet Data Access Request Form 13

8. Reference List 15

# Overview

H3ABioNet has been tasked with securely storing research data belonging to the H3Africa consortium. The data will be stored at the Central Node, which is the Computational Biology (CBIO) Group located at the University of Cape Town’s Medical Campus.

In addition to securely storing the H3Africa research data, H3ABioNet will facilitate access to this data as instructed by the H3Africa Consortium policies, and will facilitate submission of the data to the European Genome-Phenome Archive (EGA). Data should be submitted to H3ABioNet after it has been through the relevant quality control processes.

# Purpose

This policy document details the data storage and access best practices put in place to secure the H3Africa data while stored at H3ABioNet and defines the preferred methods of data submission and data access.

# Data Submission Policy

H3ABioNet does not own any of the H3Africa research data and as such will not verify or interrogate any data submitted, unless specifically requested to do so by the PI who owns the data.

A data submission request (DSR) downloadable from [www.h3africa.org](http://www.h3africa.org) needs to be completed by the H3Africa Project Principal Investigator (PI) and submitted in writing to the H3ABioNet Data Team (HDT) at [h3adata@cbio.uct.ac.za](mailto:h3adata@cbio.uct.ac.za) notifying them of an impending data delivery. The email subject needs to state the following: *H3Africa Data Submission*. We strongly recommend providing at least 1 months’ notice of an impending submission so that storage space can be provisioned.

Anonymization…

When submitting the actual data, a high-level folder needs to be created using the following naming convention:

PI\_ShortDate\_Country\_SubmissionNumber  
e.g. NicolaMulder\_27Sep13\_SA\_00 | 00 would increment by one for each submission

All supporting files will be included within this high-level folder. The following supporting files / documentation are expected at a minimum:

* All data files
* A readme file listing all the files in the submission and a brief description of these files
* Example of a blank consent form
* Description of any restrictions on the usage of the data
* Phenotype description files

The DSR request is to be sent by the DAC to the HDT mailing list which will contain the following members:

* + - Suresh Maslamoney | [suresh.maslamoney@uct.ac.za](mailto:suresh.maslamoney@uct.ac.za)
    - Ayton Meintjes | [ayton@cbio.uct.ac.za](mailto:ayton@cbio.uct.ac.za)
    - Gerrit Botha | [gerrit@cbio.uct.ac.za](mailto:gerrit@cbio.uct.ac.za)
    - Sumir Panji | [sumir.panji@uct.ac.za](mailto:sumir.panji@uct.ac.za)
    - Nicola Mulder | [nicola.mulder@uct.ac.za](mailto:nicola.mulder@uct.ac.za)

The HDT will contact the individual wanting to submit a dataset to determine the best method to be used for the submission of the data.

The DSR will be documented and archived –the applicable data transportation workflow will then be followed (see point 4 below).

NB. The purpose of the H3ABioNet archive is to allow controlled access to the data by other consortium members, and to prepare it for submission to EGA. It is not intended as a primary datastore and PIs are advised to maintain their own copy of the files at their site.

The H3Africa internal website will maintain a list of:

* Projects for which data has been submitted
* Description of this data
* Dates of submission, and
* PI contact details

## 1.1 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/>). To ensure that files can be recognised and traced to the correct project, please use the following naming convention:

PI\_ShortDate\_Country\_SubmissionNumber  
e.g. NicolaMulder\_27Sep13\_SA\_00 | 00 would increment by one for each submission

**The minimum file requirements for EGA sequence data summarized as per the EGA guidelines are:**

*A more detailed list of the EGA requirements is accessible at* [*https://www.ebi.ac.uk/ega/*](https://www.ebi.ac.uk/ega/)

* Each run is submitted with files containing data for a single sample only
* Samples submitted to the EGA should be accompanied with information regarding gender, donor ID (anonymised individual identifier) and phenotype information critical for facilitating analysis (for example, defining your tumour and non-tumour samples and/or defining disease state using controlled ontology terms)
* Gender should be described as 'male', 'female' or 'unknown'. If 'unknown' due to a known sex chromosome aneuploidy, please create a user defined attribute called 'Sex chromosome karyotype' and add the appropriate value, for example, 'XXY'
* The EGA recommends using the [Experimental Factor Ontology Database](http://bioportal.bioontology.org/ontologies/50419?p=terms) for describing your sample phenotypes.

**The suggested file formats based on EGA requirements are:**

|  |  |
| --- | --- |
| **Data Format** | **Description** |
| Array type data | The EGA accepts processed data from all types of array based technologies, such as genotypes, gene expression, methylations, etc.  The EGA also archives any associated phenotype information. |
| GWAS data | The EGA accepts submissions of complete summary level data associated with processed data such as genotypes, structural variants and whole genome sequence with any value associated with these calls. As an example, summary level data associated with genotypes called with separate algorithms can be submitted.  **If applicable, please ensure that your submission of summary level data does not contravene the original consent agreements signed by the participants of the study.**  **WE DO NOT ACCEPT SUMMARY SUBMISSIONS BASED ON TOP LEVEL SNPS.** |
| NGS data | * The minimum submission level for EGA is base/colour calls with quality scores (Signal data is no longer accepted for Illumina GA/Hiseq and SOLiD platforms but continues to be supported for the 454 platform). * BAM format is the preferred EGA data submission format. All EGA submitters are encouraged to submit their data using BAM files. * All submitted BAM files must be readable with [SAMtools](http://samtools.sourceforge.net/) and [Picard](http://picard.sourceforge.net/). Currently, BAM files must be de-multiplexed prior submission. However, we plan to shortly accept submissions of BAM files containing reads from multiple samples (**NB Colour spaced BAM is NOT supported)** * The EGA supports two different types of analyses: reference alignments and sequence variations. Reference alignments are accepted in BAM format and sequence variations in VCF format. |
| SFF Format | The recommended submission format for the 454 platform is SFF. The SFF files are nearly optimal in terms of compression and should be submitted uncompressed. |
| Illumina scarf Format | We accept but do not recommend primary data submissions in Illumina scarf format. Currently, scarf data submissions are not processed or made available in any other formats. We recommend that submitters convert their data from scarf format to Fastq format prior submission. Please note, that scarf format typically uses log-odds qualities that should be converted into Phred qualities when preparing the Fastq files. If submitted, scarf files should be compressed using gzip or bzip2. |
| Illumina qseq Format | We accept but do not recommend primary data submissions in Illumina qseq format. Currently, qseq data submissions are not processed or made available in any other formats. We recommend that submitters convert their data from qseq format to Fastq format prior submission. If submitted, qseq files should be compressed using gzip or bzip2. |
| Fastq Format | Primary sequence data submissions of single and paired reads are accepted as [Fastq](http://en.wikipedia.org/wiki/FASTQ_format#Quality) files that meet the following the requirements:   * Quality scores must be in [Phred](http://en.wikipedia.org/wiki/Phred_quality_score) scale. For example, quality scores from early Solexa pipelines must be converted to use this scale. Both ASCII and space delimitered 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 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. |

**Note**: Not all the required files will necessarily be available at this point in the analysis, as submission is due after quality control. Additional analysis results files can be submitted at any stage before submission to EGA.

**Additional information required at time of submission**

Information on the submitting project:

* PI name
* PI email address
* Project title
* Description of dataset
* Restrictions on use of data
* List of names of people on project who should have unrestricted access to the data
* List of names of people who can have limited access to the data.

# Data Access Policy for H3Africa Consortium Members

H3ABioNet will not knowingly distribute H3Africa research data to anyone without the express written approval of the PI who owns the data.

* 1. Once data is received, a description of the data along with the contact details of the relevant data owner will be added to the H3ABioNet website.
  2. To access any data, individuals need to request access via the relevant data owner as stipulated on the H3ABioNet website.
  3. Once a data access request (DAR) has been approved by the relevant PI, confirmation will be submitted in writing to the HDT granting access of a subset of data to a specific individual.
  4. The DAR request will be sent to the HDT mailing list which will contain the following members:
     1. Suresh Maslamoney | [suresh.maslamoney@uct.ac.za](mailto:suresh.maslamoney@uct.ac.za)
     2. Ayton Meintjes | [ayton@cbio.uct.ac.za](mailto:Ayton@cbio.uct.ac.za)
     3. Gerrit Botha | [gerrit@cbio.uct.ac.za](mailto:Gerrit@cbio.uct.ac.za)
     4. Sumir Panji | [sumir.panji@uct.ac.za](mailto:sumir.panji@uct.ac.za)
     5. Nicola Mulder | [nicola.mulder@uct.ac.za](mailto:Nicola.mulder@uct.ac.za)
  5. The HDT will contact the individual requesting the data and determine the best method of data access as documented in point 3 above –the applicable workflow will be followed.

# Data Storage

Data sent to H3ABioNet will need to adhere to the Data Submission Requirement (DSR) as documented above. This data will be stored on the H3ABioNet Central (CBIO) central Storage Area Network (SAN) array and have the following data security best practices imposed on it.

**Note:** Data access security measures employed can and will change based on regularly reviewing the currently employed security methods. CBIO reserves the right to change these data access security measures without notice.

## 3.1. Security Best Practices

* + 1. A SAN device will be used to physically store the H3Africa data.
    2. This SAN device is located in the CBIO server room which is protected by swipe card access in a lockable network cabinet.
    3. Access to this room is limited to IT and maintenance personnel only.
    4. The server managing this data will not be internet facing.
    5. When granting access to data, a user account will be created and kept live for 24 hours (where possible) -see data transfer section 3 below.
    6. When requesting data, the requested data will be moved to a secure FTP server for access and recorded as follows:
       1. Who the individual being granted access to the data is.
       2. The relevant documentation granting the individual access.
       3. What data was requested.
       4. The date and time the user account was created and disabled will be recorded and archived.

## Disaster Recovery

The CBIO branch of H3ABioNet where the physical data will be stored employs a disaster recovery (DR) infrastructure. All H3Africa research data stored at CBIO will automatically be incorporated into this DR infrastructure listed below:

**Note:** The DR infrastructure employed by CBIO can be changed without notice.

* + 1. The SAN which holds the physical H3Africa data has been designed with internal redundancy for automatic failover:
       1. Two power supplies connected to two separate power inlets.
       2. Inline UPS power for immediate failover power and an external generator backup power for longer outages.

**Note:** *The UPS and generator backup power does not guarantee access to data during a power failure; these measures were put in place to ensure a safe shut down of the network infrastructure hardware to limit data loss and does not take into account internet facing equipment.*

* + - 1. Dual network cards for load balancing and failover.
      2. Dual iSCSI controller cards for load balancing and failover.
    1. Data is backed up to physical tape media in accordance with the CBIO backup schedule. In summary, CBIO does monthly full backups, weekly incremental backups and daily differential backups.
    2. Data on the physical tape media is stored off-site and kept for three months, four weeks and one week respectively.
    3. The SAN device carries a 5 year next business day supplier’s on-site warranty to limit downtime.
    4. The server room employs a FM200 gas infrastructure in the event of fire.
    5. The server room employs a climate control monitoring and notification system.

# Data Transfer

This applies to the submission of data to H3ABioNet as well as transfer of data from H3ABioNet to H3Africa Consortium members.

Internet bandwidth is often quite expensive, slow and at times, unreliable on the African continent. It is with this in mind that H3ABioNet will employ two types of physical data transportation methods. When requesting or submitting small datasets, a small dataset is considered anything smaller than 5GB –a secure FTP website can be used. For those file transfers larger than 5GB or in areas where internet is slow or unreliable, we recommend using the courier method of data access / submission. The two methods are further defined below:

## 4.1. Online Workflow Method

* + 1. Data will be made available to individuals via a secure FTP website when requesting smaller datasets. Anything 5GB or smaller in size is considered to be a smaller dataset.
    2. Request by an H3Africa Consortium member to access a subset of data should be arranged directly with the PI who owns the data. Confirmation of access should then be sent to the HDT mailing list.
    3. CBIO will make the requested subset of data available to this individual via a secure FTP website.
    4. CBIO will create a temporary account on their secure FTP website which will be available for 24 hours for the individual based on the access request.
    5. Access will be recorded and attached to the original data access documentation received in point 2 above.

## 4.2. Courier Workflow Method

* + 1. When requesting larger datasets, it is recommended to use the courier method to transfer data.
    2. Requests by an H3Africa Consortium member to access a subset of data should be arranged directly with the PI who owns the data. Confirmation of access should then be sent to the HDT mailing list.
    3. Based on confirmation of granted access, the HDT will contact the respective individual using the contact details listed in the DAR received.
    4. CBIO will commission a courier company to deliver the requested data on physical hard drives to the individual.
    5. The individual would need to move the data from this physical hard drive and return the hardware to the CBIO offices in Cape Town in a timely manner.

# Responsibilities

H3ABioNet will in no way own the H3Africa data. H3ABioNet acts only as a conduit to securely store the H3Africa data and make subsets of this data available to individuals based on DARs received from the H3Africa PI who owns the data.

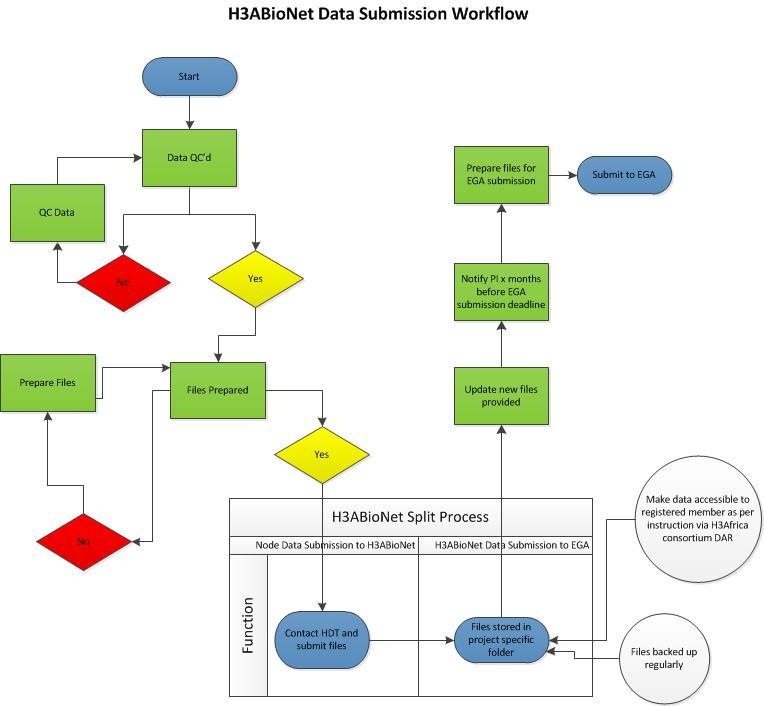
* 1. H3ABioNet will not be responsible for vetting DARs. H3ABioNet will only act on the requests approved by the relevant H3Africa PI who owns the data.
  2. H3ABioNet will not be responsible for ensuring that the requesting individual adheres to storage and usage as per the applicable federal, state, and local laws or participant consent.
  3. H3ABioNet will not be responsible for ensuring the validity of the data.
  4. H3ABioNet will not make the final decision when granting access to this data.
  5. H3ABioNet will not grant access to this data to any individual without the express consent of the Project PI.

# Resources

|  |  |
| --- | --- |
| **Resource** | **Description** |
| h3adata@cbio.uct.ac.za | Mailing list used for data submission’s and access requests |
| 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 |
|  |  |

# Appendix

## Data Submission Workflow



## H3ABioNet Data Submission Form

|  |  |
| --- | --- |
| DSR001 | **H3ABioNet Data Submission Form** |

###### Notes:

* Data Submission Request forms can be downloaded from the [www.h3africa.org](http://www.h3africa.org) and [www.h3abionet.org](http://www.h3abionet.org)
* This form is to be used by PI’s wishing to submit data for storage at the CBIO node of H3AbioNet
* All fields should be completed in full
* The completed form should be emailed to [h3adata@cbio.uct.ac.za](mailto:h3adata@cbio.uct.ac.za)
* Email subject should be: *H3ABioNet Data Submission Request*
* Adhere to the data naming convention: High-level folder named as follows: PIName\_ShortDate\_Country\_SubmissionNumber

E.g. ***NicolaMulder\_27Sep13\_SA\_00*** | *00 would increment by one for each submission*

* All supporting documents are to be contained within this single high-level folder
* Folder to include a readme file listing all files contained in submission and a brief description of files
* The terms and conditions statement needs to be accepted to validate this request

###### Terms and conditions

I/We confirm that this data to be submitted has been granted ethical approval and is consistent with the informed consent of the participants to upload data to the H3ABioNet secure storage for archiving purposes and complies with the applicable laws and regulations of the originating country. Once submitted, access to this data will only be made available based on the approval of the data owner and or the H3Africa DAC by way of a completed DAR001 form.

Should there be any change to the ethical approval, laws or regulatory conditions applicable to this data submission. I/we the data owner/s will ensure that H3ABioNet receives the relevant updated information.

I/we the data owner/s accept these terms and conditions:  Yes  No

###### Principle Investigator / Data Owner Details

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PI / Data Owner:** | | |  | | |
| **Node:** |  | | | | |
| **Office Number:** | |  | | **Alternative Number:** |  |
| **Email Address:** | |  | | | |
| **Physical address:** | | |  | | |
| **Postal Address:** | |  | | | |

###### Data Details

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Project Title:** |  | | | | |
| **Data File Type:** | | Fastq Illuminate qesq Illuminate scarf SFF BAM VCF bzip2 gzip  other  If other, please specify:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | | |
| **Brief Description of Data:** | | |  | | |
| **Has a high-level folder with the correct naming convention been created?** | | | | Yes  No | |
| **Does the high-level folder include a readme file listing all files included in the submission with a brief description of these files?** | | | | | Yes  No |
| **List of Names of People who should have unrestricted access to the data:** | |  | | | |
| **List of names of people who can have limited access to data:** | |  | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PI Signature:** |  | | **Date signed:** |  |
| **H3ABioNet Signature:** | |  | **Date Signed:** |  |

###### For Office Use Only

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Name of Capture:** |  | | **Date DSR Received:** | | |  |
| **Has all required information been provided on this form?** | | | Yes  No | | | |
| **Has this completed form been signed off by the PI / data owner?** | | | | | Yes  No | |
| **Does the data file have a valid name format?** | | Yes  No | | | | |
| **Has this request been captured on the H3ABioNet database?** | | | | Yes  No | | |
| **Has a folder with the necessary access controls been created on the SAN array?** | | | | | | Yes  No |
| **Has this form been filed in the H3ABioNet data storage file?** | | | | Yes  No | | |

**Additional Notes**

|  |
| --- |
|  |

## 7.3. H3ABioNet Data Access Request Form

|  |  |
| --- | --- |
| DAR001 | **H3ABioNet Data Access Request Form** |

Notes:

* Data Access Request forms can be downloaded from the [www.h3africa.org](http://www.h3africa.org) and [www.h3abionet.org](http://www.h3abionet.org)
* This form is to be used by PI’s wishing to gain or grant access to their stored data to individuals
* All fields should be completed in full
* The completed form should be emailed to [hdt@cbio.uct.ac.za](mailto:hdt@cbio.uct.ac.za)
* Email subject should be: *H3ABioNet Data Access Request*
* *The terms and conditions statement needs to be accepted to validate this request*

**Terms and Conditions**

I/we the data owner/s confirm that the information supplied in this form is correct and true. Access to any H3Africa data archived at the CBIO node of H3ABioNet will only be granted to the specified individual listed in the below section “Individual requesting access details”, once a completed DAR001 form has been received from the data owner/s or the H3Africa DAC.

Data will be made available either via FTP for download or physical hard drive couriered to the requester. Data transfer /access methods are further defined in the H3ABioNet data access policy downloadable from [www.h3abionet.org](http://www.h3abionet.org) or [www.h3africa.org](http://www.h3africa.org)

Where FTP access is requested, the data owner/s need to specify how long this data needs to be made available for. A temporary login account will be created based on this life span.

H3ABioNet is **NOT** responsible for ensuring that the requester conforms to the laws, regulatory and ethical approval applicable to this data.

I/we the data owner/s accept these terms and conditions:  Yes  No

**Principle Investigator / Data Owner Details**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **PI / Data Owner / DAC Representative:** | | |  | | | | |
| **Node:** |  | | | | **Country:** | |  |
| **Office Number:** | |  | | **Alternative Number:** | |  | |
| **Email Address:** | |  | | | | | |
| **Physical address:** | | |  | | | | |
| **Postal Address:** | |  | | | | | |

**Details of data to be accessed**

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Title:** |  | | |
| **High-level Folder Name:** | |  | |
| **Type of access required:** | | Online access Physical copy sent to individual  For online access, how long should this data be available:\_\_\_\_\_\_\_\_\_\_\_\_ | |
| **What portion of the data is to be made available?** | | |  |
| **Proposed research use of requested data** | | |  |

**Individual requesting access details**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Full Name:** | | |  | | | | |
| **Node:** |  | | | | **Country:** | |  |
| **Office Number:** | |  | | **Alternative Number:** | |  | |
| **Email Address:** | |  | | | | | |
| **Physical address:** | | |  | | | | |
| **Postal Address:** | |  | | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PI Signature:** |  | | **Date Signed:** |  |
| **H3Africa DAC Representative Signature:** |  | | **Date Signed:** |  |
| **H3ABioNet Signature:** | |  | **Date Signed:** |  |

**For Office Use Only**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name of Capture:** |  | **Date DAR Received:** | | |  |
| **Has all required information been provided on this form?** | | Yes  No | | | |
| **Has this completed form been signed off by the PI and H3Africa DAC?** | | | | Yes  No | |
| **Has this request been captured on the H3ABioNet database?** | | | Yes  No | | |
| **Has this form been filed in the H3ABioNet data storage file?** | | | Yes  No | | |

**Additional Notes**

|  |
| --- |
|  |

# Reference List

* 1. EGA - <https://www.ebi.ac.uk/ega/>