H3AFRICA CONSORTIUM DATA SHARING, ACCESS AND RELEASE POLICY

Last updated April 2020

Table of Contents

Background

The goal of the Human Heredity and Health in Africa (H3Africa) Initiative (http://h3africa.org/) is to enhance the capacity of African researchers to conduct state of the art genomics research. H3Africa aims to undertake cutting edge research to advance understanding of the genetic and environmental determinants of common diseases in Africa and to use this knowledge to improve the health of African populations. The H3Africa data sharing, access and release policy is built upon H3Africa principles of ethics, governance and resource sharing that have been established by the NIH and Wellcome Trust (http://h3africa.org/ethics_governance_resourcesharing.cfm) and the H3Africa consortium members. These principles aim to strike an appropriate balance in ensuring that adequate safeguards are in place to protect participants, while maximizing the ability of investigators to advance research.

Principles for the H3Africa Initiative data sharing, access and release policy include:

- Maximizing the availability of research data, in a timely and responsible manner.
- Protecting the rights and privacy of human subjects who participated in research studies.
- Recognizing the scientific contribution of researchers who generated the data.
- Considering the nature and ethical aspects of proposed research whilst ensuring the timely release of data
- Promoting deposition of genomic data in existing community data repositories whenever possible.

General Guidelines

The H3Africa Initiative is committed to providing research data generated by the H3Africa research projects to the entire research community. Data to be released includes genomic and phenotypic data from appropriately consented individuals. After an initial embargo period to enable H3Africa projects to conduct analysis and share data amongst projects, as per international norms for large data generating consortiums, the data will be released to the wider scientific community under controlled access. In compliance with current international standards to protect participant confidentiality, the H3Africa-generated genotype and phenotype data will be available to *bona fide* researchers within the scientific community through a controlled access process at the European Genome-phenome Archive (EGA). Where applicable, e.g. for non-human data, H3Africa projects may deposit their sequence data at the European Nucleotide Archive (ENA), which will be publicly accessible to all. Genome summary results (GSRs) may be made available separately either under controlled or open access depending on the conditions of the study. H3Africa microbiome or other metagenomics data will be deposited in the ENA, but since this is an openly accessible resource, any human sequence contamination will need to be removed prior to submission and submitted to the EGA together with the relevant phenotype information. The phenotype information for these datasets will then be linked to their ENA entries for the non-human microbiome data and will be available via controlled access. To gain access to H3Africa datasets in the EGA, researchers must submit a request to use the data by registering on the H3Africa catalogue website and submitting the online request form. The data access request application is reviewed and approved by an independent H3Africa Data and Biospecimen

Access Committee (DBAC). All submitted data access requests must be accompanied by a signed Data Access Agreement (DAA) which lays out the terms and conditions under which access to H3Africa data will be granted.

Specific Guidelines

Data Sharing and Release

Data types to be generated by the H3Africa consortium may include some of the following data elements:

Phenotype data

- Demographic information
    - Age
    - Sex
    - Participant's (Country of birth; native language; tribal affiliation)
    - Father's and Mother's (Countries of birth; native languages; tribal affiliations)
    - Socio economic factors (highest level of schooling; employment status)
- Anthropometric data
    - Weight
    - Height
    - Blood pressure readings
- Disease and health related phenotype data
    - Smoking status
    - Alcohol use
    - Drug use
    - Kidney disease related
    - Cardiovascular disease related
    - Stroke history
    - HIV related
    - Cancer related
    - Sickle Cell disease related
    - Dyslipidaemia related
    - Infectious disease related
    - Mental health related
    - Chronic conditions
- Environmental exposures data

Genetic Variation data from both human and pathogen (whole-genome or exome sequence data, genotyping data from arrays, any other omics data being generated)

- Sequence data
    - Primary data following QC
        - FastQ files linked with de-identified participant ID (minus adapters, linkers, barcodes)
        - Binary Alignment files (BAMs, de-multiplexed) – linked with de-identified participant ID
    - Variant calls following QC

        - Variant calling files (VCFs)
- Epigenetic data, e.g. DNA methylation
- Genotype data from arrays
    - Raw intensity files linked with de-identified participant IDs (IDATs, CELs)
    - Primary array data following QC
    - Variant calls following QC

Microbiome sequence data

- Patient/sample phenotypes
- 16S rRNA sequence data for microbiome
    - FastQ files linked with de-identified participant ID (minus adapters, linkers, barcodes)
    - Final analyses BIOM files (e.g. OTUs)
- Non-human full genome sequence or proteomic data for microbiome
- Possible human sequence contamination
- Metaproteomics data from mass spectrometry

Additional information for a dataset includes:

- Study type e.g. case / control, trio, longitudinal
- Technology platform
- Mapping file indicating relationship between files

The process and timelines are shown in figure 1.

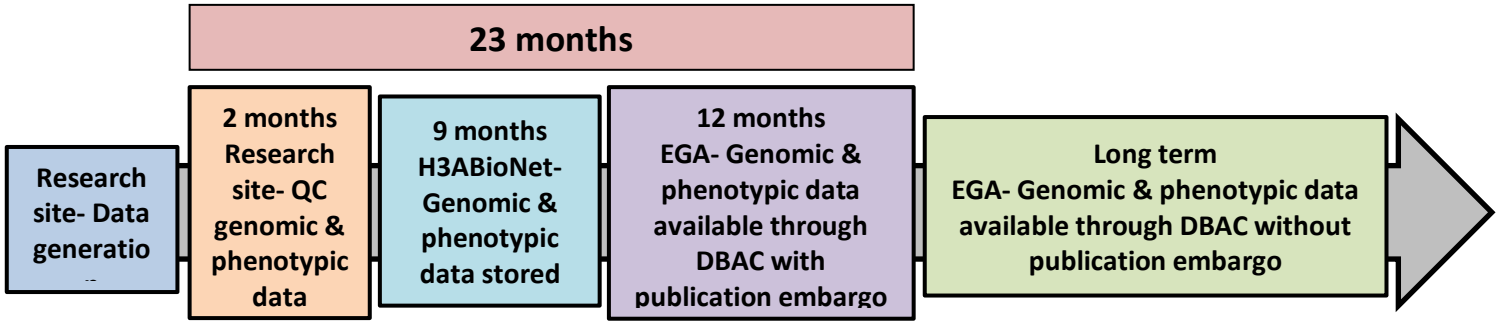

Figure 1. Proposed timeline for data submission and access for H3Africa.

The Wellcome Trust and NIH data sharing policies can be found here: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Data-management-and-sharing/index.htm and https://grants.nih.gov/grants/policy/data_sharing/

Whole genome, microbiome or targeted (exome or regional capture) sequence data and DNA sequence variants that have been quality controlled and validated, will be submitted to the H3Africa Archive at H3ABioNet within approximately 2 months of data acquisition from the data generating facility (Note, it is understood that large datasets may take longer to QC, and therefore a motivation can be submitted for an extension. The PIs should discuss such issues

with their grant officer and come to an agreement on a submission date). In some cases, additional computation is required prior to QC, such as imputation. The data should then be submitted once the dataset is QC'ed and ready for analysis and interpretation. What constitutes a complete dataset for submission and the actual timelines and dates of deposition of the Quality Controlled (QC) data to H3ABioNet will be agreed upon between the H3Africa project PI and the appropriate funding agency. See Data Quality Control and Validation Guidelines (Appendix A) and H3ABioNet Data Submission Guidelines (Appendix B) for more information on the QC process and the process for submission of data to H3ABioNet.

Phenotypic data will be submitted at the same time as submission of the quality controlled genomic data to H3ABioNet. H3Africa projects are encouraged, whenever possible, to collect and submit a quality controlled and validated set of recommended core phenotypic data fields as defined by the Phenotype Harmonization Working Group. A data submission template containing the required metadata for EGA data submission and Phenotype Harmonization Working Group recommended phenotypic data fields will be provided to consortium members. Microbiome data should also be submitted to H3ABioNet and the same timelines apply even though the sequence data will be submitted to the ENA rather than EGA.

Quality controlled genomic data and phenotypic data will be held in H3ABioNet for nine months. During this nine-month period:

- H3ABioNet will help the PIs to format the data in preparation for submission to EGA. See EGA Data Submission Guidelines (Appendix C).
- H3Africa data generators can analyse and publish their datasets and analysis.
- H3Africa data generators should submit updates of their data to H3ABioNet – this includes new batches of data as well as updates to data already submitted e.g. new variant calls, regenerated BAM files, extended phenotypic data, etc.
- H3Africa consortium members can request access to the data from the H3Africa data generators. Collaborative plans should be discussed and the intent to publish should be submitted using the Manuscript Concept Document as explained in the H3Africa Publications Policy.

Following the nine-month data hold at H3ABioNet, data will be sent to and publicly released through the EGA.

- Datasets deposited into EGA will become publicly available to bona fide[1] scientists who request access through an independent H3Africa DBAC with a twelve-month publication embargo on the dataset. H3Africa Consortium members will access data through EGA after the 9 months hold at H3ABioNet, unless they enter a direct collaboration with the data owner. See Appendix D. Data and Biospecimen Access Committee Guidelines and Appendix F. Data Access Agreement.

Publication Embargo

To protect the publication rights of the investigators who generate data in H3Africa, a publication embargo will be used. The embargo, which will be managed by the H3Africa DBAC secretariat, will enable grant investigators exclusive rights to publish analyses of the dataset for one year following the release of genotype or sequence datasets through the EGA, including the raw, pre-analyzed data. H3Africa has defined the aims and scope of the projects and these are outlined on the H3Africa website (http://www.h3africa.org). During this period of exclusivity, access to the data will be granted to other investigators through the DBAC, who may analyze the data, but are expected not to submit their analyses or conclusions for publication during the exclusivity period. Publication exclusivity will extend to all forms of public disclosure, including meeting abstracts, oral presentations, and publicly accessible electronic submissions (e.g. websites, web blogs). During this time period PIs should consider pre-publication prints to BioRxiv.

Data will no longer be subject to the publication embargo if the dataset (a dataset in EGA can include genotype or phenotype data or both and a project may submit more than one dataset) has been published in a peer reviewed journal, or if one year has passed since the dataset used for analysis was made available through the EGA. Each project is treated individually so that an embargo lifted due to publication of one H3Africa project does not affect the ongoing embargo on other H3Africa projects. In addition, dates of embargoes will differ between datasets within and between the projects depending on their dates of submission of data to EGA. Any publication on a specific dataset lifts the embargo on that dataset, however, if one publishes on a portion of the data, only that data becomes available –this applies to datasets that are publishable in sets.

Details of the publication embargo dates will be made available to the H3Africa DBAC Secretariat along with the list of datasets accessed from the EGA. Potential data users will be encouraged to contact the H3Africa DBAC Secretariat if there is any doubt about how the data should be used. The DBAC secretariat contact details will be made available for questions.

It is expected that investigators who access H3Africa data acknowledge the Consortium and relevant project(s) appropriately in any oral or written presentations, disclosures, or publications of the analyses.

European Genome-Phenome Archive (EGA)

This information was obtained from the EGA website (https://www.ebi.ac.uk/ega/). The EGA is a service for permanent archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects. Information on EGA submission guidelines is provided in Appendix C. The EGA contains exclusively data collected from individuals whose consent agreements authorize data release only for specific research use or to bona fide researchers. Strict protocols govern how information is managed, stored and distributed by the EGA project. Once processed, all data are encrypted for dissemination and the encryption keys are delivered offline.

H3ABioNet will serve as the data coordinating centre that works directly with the EGA for submission of all H3Africa data on behalf of the H3Africa projects. H3Africa projects with data in the Archive will be kept informed of the progress and status of data submission to the EGA.

The EGA will create new or update existing accounts only from the direction of the H3Africa DBAC. The EGA will also require authorization from the H3Africa DBAC when the original application (access request) must be updated, as access is granted for a fixed period.

Approved data users are provided with an EGA account that includes details of the applicant, such as the name, postal address, email and it lists all accession rights granted by the different consortia and DBAC. The account exists solely within the EGA system. The EGA account is for personal use only; the terms and conditions of the account prohibit sharing of the account details. The users will be able to request encryption keys, access authorized data files or request FTP/Aspera accounts from the system by using the tools linked to the account.

Data Access

The basic descriptive information about each H3Africa study will be available publicly through the EGA and the H3Africa Catalogue. Access to the genotype and phenotype datasets submitted and stored in EGA will be provided to data users approved by the H3Africa DBAC.

Potential data users will be required to provide a summary of their proposed research (see Data Access Request Form in Appendix E). The review of requests for data access will include screening for inappropriate use of the data and to ensure that it complies with the ethical consent associated with the data or samples. Access decisions will be based on proposed area of research, not on scientific strength or novelty. There will be no scientific-peer review of the proposed research and it is accepted that duplicative analyses may occur and will not be excluded.

Approved data users will be required to agree to the Terms and Conditions of a Data Access Agreement (DAA), which aims to protect the privacy and interests of the research participants. The DAA will require users to agree to:

a) Use the data only for the approved research;
b) Protect participant confidentiality;
c) Follow appropriate data security protections;
d) Follow all applicable laws, regulations and local institutional policies and procedures for handling genotype, sequence and phenotype data;
e) Not attempt to identify individual participants from whom data within a dataset were obtained;
f) Not sell any of the data elements from datasets obtained from the EGA;
g) Not share with individuals other than those listed in the request any of the data elements from datasets obtained from the EGA;
h) Point out any problems or concerns about the data to EGA who will then pass it on CC or PI
i) Agree to the listing of a summary of approved research uses within the EGA along with his or her name and organizational affiliation;
j) Agree to report, in real time, violations of the H3Africa Data Access policy to the DBAC;
k) Acknowledge the H3Africa policy with regard to publication and intellectual property;

l) Respect first publication rights as described in the embargo;
m) Cite the original H3Africa publication that describes the analysis of the dataset(s) used, and;
n) Provide annual progress reports and publications on research using the genotype, sequence and or/phenotype dataset. These will be collected by the H3Africa DBAC Secretariat.

A list of projects for which access has been approved (PI and title) will be made publicly available on the H3Africa website. The H3Africa DBAC Secretariat will keep records of who gained access, and when, to ensure that the terms of access and the conditions of the publication embargo are complied with. These records will not be made publicly available but will be available to the H3Africa Steering Committee for oversight purposes.

When external users request access to H3Africa biospecimens, they will be required to submit any new data generated from those samples to the EGA under the H3Africa project title.

Data and Biospecimen Access Committee

Data access requests will be managed through an independent Data and Biospecimen Access Committee (DBAC) by appointed members with relevant expertise in areas such as the relevant particular scientific disciplines, research participant protection, and privacy (Appendix D). Investigators and institutions seeking data from the EGA will be expected to meet data security measures (such as physical security, information technology security, and user training) and will be asked to submit a data access agreement, that is co-signed by the investigator and the designated Institutional Official(s). Data access requests should include a brief description of the proposed research use of the requested dataset(s) (see Appendix E). The DBAC will be responsible for receiving and reviewing applications for access to H3Africa data and determining whether:

- The principal investigator is a bona fide researcher;
- The research use is acceptable;
- The DAA has been completed satisfactorily and includes all required signatures.

Bona fide scientists working for commercial companies may request access to datasets for research or “commercial” purposes if the consent allows for this. Since consent forms differ, such access will be reviewed on a case-by-case basis. The DBAC or their designees will review requests for access to determine whether the proposed use of the dataset is scientifically and ethically appropriate and does not conflict with constraints or informed consent limitations identified by the Investigators that submitted the dataset to the EGA. In the event that requests raise concerns related to privacy and confidentiality, risks to populations or groups, or other concerns, the DBAC will consult with other experts as appropriate.

Release of patient/donor identifying data:
H3Africa may release curated and coded phenotype, exposure, genotype, and pedigree data. These detailed data will be made available through a controlled access process through the EGA. To protect the rights and privacy of human subjects who participate in the studies, clinical metadata, genomic, or other datasets, or a subset of the clinical and other metadata that may potentially identify the human subjects who donated samples **shall not** be released in any publicly accessible databases. Clinical data and other fields that may potentially uniquely identify an individual will be carefully reviewed and flagged prior to sharing and releasing to publicly accessible databases.

In order to minimize risks to study participants and their communities, data submitted to the EGA will be de-identified and submitted using anonymous participant identifiers assigned by the projects to ensure the identities of data subjects cannot be readily ascertained or otherwise associated with the data by the repository staff or secondary data users. Further details are provided in the EGA Submission Guidelines (Appendix C).

Monitoring
A breach of any of the conditions of the Data Access Agreement (DAA), including failure to respect the publication embargo, will terminate the DAA and access to the data will be withdrawn. Future access may also be denied to individuals found responsible for a previous breach of the conditions of the DAA. In addition, when a breach of the publication embargo is suspected or takes place, the H3Africa Steering Committee and/or the DBAC may contact the appropriate journal editor with evidence that data use conditions have been breached and to request that any manuscripts be withdrawn.

Data users will be expected to submit annual reports to the DBAC Secretariat describing the use of the data to ensure that it complies with the terms of the DAA and the provisions laid out in the associated informed consent. These will be requested and monitored by the H3Africa DBAC Secretariat.

Intellectual Property

It is the aim of H3Africa that genotype-phenotype data made available through the EGA and all conclusions derived directly from them will remain freely available, without any licensing requirements, for uses such as, but not necessarily limited to, markers for developing assays and guides for identifying new potential targets for drugs, therapeutics, and diagnostics. H3Africa discourages any premature claims on pre-competitive information that may impede research, though it encourages patenting of technology suitable for subsequent private investment that may lead to the development of products that address healthcare needs.

The filing of patent applications and/or the enforcement of resultant patents in a manner that might restrict use of H3Africa genotype-phenotype data could diminish the potential public benefit they could provide. Approved users and their institutions, through the execution of a Data Access Agreement, will acknowledge the goal of ensuring the greatest possible public benefit from H3Africa data.

NIH Intellectual Property Policy

By requesting access to genomic dataset(s), the Requester and Approved Users acknowledge the intent of the NIH that anyone authorized for research access through the attached Data Access Request follow the intellectual property principles within the NIH Genomic Data Sharing Policy (https://osp.od.nih.gov/wp-content/uploads/NIH_GDS_Policy.pdf) as summarized below~~:~~

NIH encourages patenting of technology suitable for subsequent private investment that may lead to the development of products that address public needs without impeding research. However, it is important to note that naturally occurring DNA sequences are not patentable in the United States.~~51~~ Therefore, basic sequence data and certain related information (e.g., genotypes, haplotypes, *p*-values, allele frequencies) are pre-competitive. Such data made available through NIH-designated data repositories, and all conclusions derived directly from them, should remain freely available, without any licensing requirements.

NIH encourages broad use of NIH-funded genomic data that is consistent with a responsible approach to management of intellectual property derived from downstream discoveries, as outlined in the NIH *Best Practices for the Licensing of Genomic Inventions* and Section 8.2.3, Sharing Research Resources, of the NIH Grants Policy Statement and found in the link provided above. NIH discourages the use of patents to prevent the use of or to block access to genomic or genotype-phenotype data developed with NIH support.

Wellcome Trust Intellectual Property policy

The Wellcome Trust Intellectual Property policy can be found here:
https://wellcome.ac.uk/funding/guidance/policy-intellectual-property

1. Definitions of a bona fide researcher:

https://www.ucl.ac.uk/epidemiology-health-care/data-sharing-faq#bonafide: bona fide researchers who:
i) conduct *bona fide* research. This involves high quality, ethical projects for research purposes using rigorous scientific methods. There must be an intention to publish the research findings for wider scientific and eventual public benefit, without restrictions and with minimal delay.

ii) have a formal relationship with a *bona fide* research organisation, which is an established academic institution, research body or organisation with the capability to lead or participate in high quality, ethical research. It is not a requirement that research is the primary business of that organisation, or that the organisation is publicly financed. In this context, a public-private partnership may qualify as a *bona fide* research organisation.

https://mrc.ukri.org/publications/browse/mrc-policy-and-guidance-on-sharing-of-research-data-from-population-and-patient-studies/: For the purposes of this guidance, key characteristics of bona fide research can be considered to be as follows: • An intention to generate new knowledge and understanding using rigorous scientific methods. (This

includes discovery research, development and validation of methodology and technology, validating and challenging previous findings, and pilot research). And... •An intention to publish the research findings and share the derived data in the scientific community, without restrictions and with minimal delay, for wider scientific and eventual public benefit. (Recognised constraints include a short prepublication delay to ensure proper management of intellectual property). And... •The intended activities are not inconsistent with legal and ethical requirements or widely recognised good research practice. In practical terms, a research project or proposal that has been approved by a recognised funder should normally be considered to be “bona fide”. A bona fide research organisation is one that has the capability to lead or participate in high quality, ethical research. It will have a public commitment to adhere to recognised research and information governance good practice. (It is not a requirement that such research is the primary business of that organisation, or that all of the research undertaken by that organisation is published. Nor is it a requirement that the organisation be publicly funded.) A bona fide researcher is a person with • the professional expertise and experience to conduct bona fide research and • a formal relationship with a bona fide research organisation that requires compliance with appropriate research governance and management systems.