H3Africa synthetic data info

The H3ABioNet synthetic data was created using a modified version of tofu (<https://github.com/spiros/tofu>), using a database of fields and values/encodings.

We used the [H3Africa Core phenotype](https://www.h3abionet.org/images/DataAndStandards/DataStandards/H3AfricaHarmonisedPhenotypes_DataDictionary_2018-05-11_Version_1.csv) data dictionary to generate new database fields and encodings to mimic African data.

## Agreed name and versioning

The synthetic data for H3Africa is named: **CINECA\_synthetic\_cohort\_Africa\_H3ABioNet\_v1**. This naming is to avoid any confusion with any real H3Africa datasets. Basic versioning will be applied to track changes both on the metadata and genetic data.

The Nextflow pipeline implemented to generate this dataset will be made available on the project Github page.

## Metadata / Phenotypic data

We used a modified version of the [TOFU](https://github.com/spiros/tofu) to generate the metadata for 100 samples. All sample identifiers are prefaced with 'fake' to avoid confusion with real datasets. We constructed a database of fields and values using the [H3Africa Core phenotype](https://www.h3abionet.org/images/DataAndStandards/DataStandards/H3AfricaHarmonisedPhenotypes_DataDictionary_2018-05-11_Version_1.csv), a set of recommended questions or variables that H3Africa should consider when designing their data collection forms. We selected a group of 47 variables out of the 255 H3Africa Core phenotypes to provide a good overlap with the CINECA core metadata model. We randomly chose categorical values from field choices from the H3Africa Core phenotype data dictionary. Continuous values, such as age and date of birth, were randomly selected from field ranges from the H3Africa Core phenotype data dictionary. For this initial version, we use the mean and median values from UK Biobank provided with TOFU to model the distribution of continuous values.

More information on the phenotypic data variables (description, data type, example) can be found in this document that describes the [coverage with the CINECA minimal metadata model](https://docs.google.com/spreadsheets/d/1Wpbf51jZjNHaIn1FJkMcMHswvHW0nasu/edit#gid=428731926).

We have started the process of normalising our variables to match the CINECA core metadata model. So far, we have harmonised a few variables, including age/birthdate, gender, ethnicity/race, height, weight, tobacco.

## Generic data

We used the [1000 Genomes project phase 3 data](http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) to generate genetic data. From the 2504 samples included in the 1000 Genomes project, we randomly selected 100 samples of African ancestries. We then used [BCFTools](http://samtools.github.io/bcftools/bcftools.html) to replace the 100 sample identifiers with the ones in the metadata file. We randomly selected 650K variants in chromosome 1 using [BCFTools](http://samtools.github.io/bcftools/bcftools.html). Because we are using a Nextflow pipeline to generate this genetic data, parameters can easily change the chromosome and variant size via the pipeline confirmation file, streamlining the whole process.

## Storage and Access

* The CINECA\_synthetic\_cohort\_Africa\_H3ABioNet1 data is currently accessible via the [CINECA synthetic data’s project google drive folder](https://drive.google.com/drive/folders/1pJoMvK-HBnMqoOTxGqkvenUVTeiQ0z8M?usp=sharing).
* We are exploring the possibility of storing this data in a storage service that offers open and authorised access to the data, together with some GA4GH APIs such as htsget, GA4GH AAI and passports for data access.
* This data will also be made available through the H3ABioNet Beacon test instance at <https://beacon2.h3abionet.org>

## Usage & Acknowledgement

Both the genetic data and metadata are fully accessible under the Creative Commons Licence (CC-BY). Any use of this data should, however, thoroughly consider the following:

* The metadata conforms to the structure and schema of the [H3Africa Core phenotype](https://www.h3abionet.org/images/DataAndStandards/DataStandards/H3AfricaHarmonisedPhenotypes_DataDictionary_2018-05-11_Version_1.csv), but it is otherwise nonsensical: no checks have been implemented across fields, and values may be completely unrealistic.
* We did not model any correlation between fields. There is, however, a plan to model a correlation on a few variables such as weight and height.
* This dataset should not be used to make any inference whatsoever as the values of the fields do not entirely reflect reality.
* Dates randomly generated are between 1910 and 1990 to avoid confusion with real data.
* This synthetic data set (with cohort “participants” / ”subjects” marked with fake) has no identifiable data and cannot be used to make any inference about H3Africa cohort data or results. This dataset aims to aid the development of technical implementations for cohort data discovery, harmonisation, access, and federated analysis. In support of FAIRness in data sharing, this dataset is made freely available under the Creative Commons Licence (CC-BY). Please ensure this preamble is included with this dataset and that the H3Africa project and the CINECA project (funding: EC H2020 grant 825775 and CIHR grant 404896) are acknowledged. If you have any questions about this dataset, please contact Mamana Mbiyavanga ([mamana.mbiyavanga@uct.ac.za](mailto:mamana.mbiyavanga@uct.ac.za)) or Nicola Mulder ([nicola.mulder@uct.ac.za](mailto:nicola.mulder@uct.ac.za)).

## 