Supplementary File S1: The working document containing items for HiGeneS Africa database and AI system for implementing the future LIMS.

Background

HI-GENES Africa database development forms part of the data management plan (DMP). The option to generate the database using one of two options.

Graph database	Relational database
A graph database is a NoSQL database that stores data as a network graph. What differentiates graph databases from other options is that they document and prioritize the relationships between data. Graph databases are made up of nodes and edges, where nodes represent specific entities, while edges represent the connection between two nodes. They are designed to be scalable and offer flexibility that's hard to find in other databases.	Relational databases store data in relational tables. Tables are defined by columns and rows, and each row is identified by a unique key so they can be linked to rows in other tables. Each individual table also includes a primary key identifying the information found within the table. For example, one table may contain customer information that relates to information in a different table containing order information. Relational databases separate the logical structures of tables and indexes from physical storage structures. This enables data professionals to make changes to physical data structure that don't affect the logical structure.

Source: https://searchdatamanagement.techtarget.com/feature/Graph-database-vs-relational-database-Key-differences

Objective: Primary objective was to discuss the process for developing user rights, entities for the entity relationship diagram and expectations for the proposed HI-GENES Africa database.

The following entities were used to build the database:

1. Instruments on RedCap
2. Consent forms (entered on RedCap)
3. Pedigrees
4. Scanned questionnaires and consent forms
5. Recruitment pictures

6. Community and public engagement material
7. LIMS data
8. Whole exome sequencing (WES) data
9. Possibly whole genome data in future
10. GWAS data still to be generated
11. Synology backup data

The database should be able to link data from all the platforms.

1. The search engine should be able to search and provide all information from all the platforms on a patient or list of patients using patient codes.
2. All project team members will have access to the platform.
    - There will be different levels of access which will depend on the need of the user to access information for data analysis, manuscript or thesis write-up or tasks as instructed by the PI.
3. Data sharing and publications must adhere to the H3Africa consortium and NIH grant conditions (please refer to the DMP).
4. The use of data obtained through the HI-GENES Africa database must be pre-approved by the PI. This relates to data analysis, scientific write-up and any collaborations with other projects.
5. Patient identifiers should not be made available to everyone except the authorized persons.
    - Who will be authorised to have access to personal identifiers?
    - Why would they need access?
        - Each site will automatically have access to personal identifiers for their site but will not have access to personal identifiers for other sites unless requested through approval from the PI.
        - There should be a process for requesting information from other sites.
        - Access to personal identifiers can be requested there is a need for follow-up visits to request additional blood samples, further information for pedigree analysis, to verify information, returning of results.

Data clean-up or validation is needed prior to the development

- There are discrepancies in the data and between sites.
- All data must be harmonised and checked before we can develop the database.

Requirements for data clean-up process

1. **Prioritise:** You will need to set aside **time** to conduct this process.
    - There were proposals for allocating a specific time each week to do this or maybe a week depending on lab activities.
2. **Coding:** Check each family and participant to ensure that sample codes are matching and correspond to the right person
    - You may not need to change old codes to new ones, the bioinformaticians will use a programming code to combine old and new codes. This is to reduce the risk of human error when relabelling tubes.
    - Ensure that the codes are the same for the same person on all data sources irrespective if it is old or new codes.
3. **Pedigrees:** Ensure that each person is on a pedigree that is well labelled.
    - The pedigree for each family must be drawn and saved with the same code as on other platforms.
4. **Complete data capturing:** Fill in all the missing data points (to the best of your ability) for each participant
5. **REDCap:** Capture the questionnaires and pedigrees on REDCap.
    - This includes qualitative data.
6. **Synology:** The data must be backed up properly in an organised manner for easy reference.
    - This includes pictures obtained during recruitment and community engagement.
7. **Sequencing data**: WES data and analysis should be linked to other platforms through patient codes. Data that will be generated in future should also be linked.