
PRACTICE PAPER

Building Infrastructure for African Human Genomic Data Management

Ziyaad Parker¹, Suresh Maslamoney¹, Ayton Meintjes¹, Gerrit Botha¹, Sumir Panji¹, Scott Hazelhurst^{2,3} and Nicola Mulder¹

¹ Computational Biology Division, Department of Integrative Biomedical Sciences, IDM, University of Cape Town, ZA

² School of Electrical and Information Engineering and Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, ZA

³ Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, ZA

Corresponding author: Ziyaad Parker (ziyaad.parker@uct.ac.za)

Human genomic data are large and complex, and require adequate infrastructure for secure storage and transfer. The NIH and The Wellcome Trust have funded multiple projects on genomic research, including the Human Heredity and Health in Africa (H3Africa) initiative, and data are required to be deposited into the public domain. The European Genome-phenome Archive (EGA) is a repository for sequence and genotype data where the data access is controlled by access committees. Access is determined by a formal application procedure for the purpose of secure storage and distribution, and must be in line with the informed consent of the study participants. H3Africa researchers based in Africa and generating their own data can benefit tremendously from the data sharing capabilities of the internet by using the appropriate technologies. The H3Africa Data Archive is an effort between the H3Africa data generating projects, H3ABioNet and the EGA to store and submit genomic data to public repositories. H3ABioNet maintains the security of the H3Africa Data Archive, ensures ethical security compliance, supports users with data submission and facilitates the data transfer. The goal is to ensure efficient data flow between researchers, the archive and the EGA or other public repositories. To comply with the H3Africa data sharing and release policy, nine months after the data is in secure storage, H3ABioNet converts the data into an XML format ready for submission to EGA. This article describes the infrastructure that has been developed for African human genomic data management.

Keywords: genomic data; data archive; h3africa data; african genomic data

1. Introduction

Advances in high throughput genomic technologies are laying the foundations for the goal of precision medicine to be realized (Christensen et al. 2015; Aronson and Rehm 2015). Decreasing costs and the capacity to generate larger volumes of human genomics data at faster rates are enabling population level genomics studies to be conducted (Goldfeder et al. 2017; Prokop et al. 2018). However, most of the current population level genomics studies and data generated to date, have a significant population representational bias with the majority of genome sequences being derived from European and North American ancestry, these are regions that have been early adopters of genomic technologies (Popejoy and Fullerton 2016; Prokop et al. 2018). African researchers, in general, have been late adopters of high-throughput technologies for use in population genomics due to more limited resources and funding. To address this critical gap in scientific knowledge about African genomics and population variation, and inspired by the African Society for Human Genetics, the National Institutes of Health (NIH) and The Wellcome Trust, through the Human Hereditary and Health in Africa (H3Africa) program, have funded multiple genomics projects led by African investigators (H3Africa Consortium et al. 2014; Mulder et al. 2018). To support the H3Africa projects in terms of provisioning of infrastructure for secure data storage, management and compute, the NIH has also funded a Pan-African Bioinformatics Network for H3Africa (H3ABioNet: <http://www.h3abionet.org>) (Mulder et al. 2016).

The H3Africa Consortium consists of multiple projects and sites distributed across Africa, most of which are generating genomic data linked to clinical data for specific diseases. The principal H3Africa funders (NIH and the Wellcome Trust) require any project data generated to be deposited into a data repository accessible by the scientific community.^{1,2} In order to facilitate the storage and accessibility of H3Africa genomics data, significant infrastructure, procedures and policies were established. Part of H3ABioNet's mandate is to develop processes and implement an infrastructure that will enable the ingestion, validation, annotation, secure storage and submission of the African genomics data to the controlled access European Genome-phenome Archive (EGA) (Lappalainen et al. 2015). This has been achieved through the development of the H3Africa Data Archive, which is also ensuring a copy of the genomic data is securely stored and retained on the African continent (Mulder et al. 2016; Mulder et al. 2017). This article describes the infrastructure that has been developed, which to our knowledge, is the first formalized human genomic data archive on the continent.

2. Methods

In order to establish the H3Africa Archive, and to make the submission process seamless, a data storage infrastructure had to be created and new processes and policies developed and adhered to.

2.1. Data Submission and Access Policy

Genomic data associated with phenotype data enables the possibility of re-identification of study participants; hence all human genomic data and its accompanying phenotypic data need to be governed by a controlled access policy (Shabani et al. 2015) (Dyke et al. 2018). H3Africa is distinguished as biospecimens are also being collected and stored at one of the three H3Africa biorepositories, so researchers can request access to genomic data and/or biospecimens. As a consortium, H3Africa has developed its own data submission and access policy which takes into account the genomics and phenotype data generated, as well as policies for the access to and transfer of biospecimens (de Vries et al. 2015). A single H3Africa Data and Biospecimen Access committee has been established to oversee the secondary use of both the data, which is being deposited in the EGA, and biospecimens in the H3Africa biorepositories. The sharing and access policies and the H3Africa Data and Biospecimen Access Committee guidelines seek to provide a balance between protecting the rights of individuals and their data, while at the same time not acting as a barrier to advancing scientific knowledge. A data requester will need to identify the data in the EGA and apply for data access (Lappalainen et al. 2015). The data access request is routed to the Data and Biospecimen Access Committee (DBAC) who review it to determine whether the intended research use is inline with the H3Africa data and access policy, and the requester is a *bona fide* researcher. Once the data request has been reviewed, the H3Africa DBAC will provide a decision to approve or reject it.

2.2. Types of data being accepted

The principal data types being collected for submission to the H3Africa Data Archive and the EGA include genomic sequence data, genotype array files, the associated phenotypes and metadata that is collected along with the samples, and results of any analysis conducted. Genomic sequence data mainly comprises of short DNA sequence reads in FASTQ format (Cock et al. 2010; Hsi-Yang Fritz et al. 2011). The types of data and associated files for the H3Africa research projects are summarised in **Table 1**.

2.3. Timelines

During participant recruitment, participants sign consent forms which gives the researcher the right to use the data for research purposes, and may or may not include consent for sharing and secondary use. Data is generated at the project sites or wherever the sequencing or genotyping equipment is located. The data then undergo validation and quality assurance by the project to clean it, which usually takes up to 2 months, though timelines vary depending on the sample size. At this point, the project's designated data submitter makes contact with the H3Africa Data Archive Team (HDAT) to begin the process of submitting their data to the archive. The submission process details are described in the results section. Once the data is accepted into the data archive's cold storage, it will be incubated for a period of 9 months giving the data owner or researcher time to analyse their data and prepare their publications (**Figure 1**). Thereafter, the data undergo final processing to ensure EGA format compliance, and then are submitted to the EGA.

¹ <https://grants.nih.gov/policy/sharing.htm>.

² <https://wellcome.ac.uk/funding/guidance/policy-data-software-materials-management-and-sharing>.

Table 1: Description of data types for submission.

| Exome/Whole Genome Sequence | 16S rRNA Microbiome studies | Genome Wide Association studies/genotyping arrays |
|--|--|--|
| Study type and description | Study type and description | Study type and description |
| Sequencing platform and technology used | Sequencing platform and technology used | Genotyping array model/name and description of the software and version used for calling the genotypes |
| FASTQ files linked with de-identified participant ID (minus technical reads such as adapters, linkers, barcodes) | FASTQ files linked with de-identified participant ID (minus technical reads such as adapters, linkers, barcodes) | Raw intensity files linked with de-identified participant IDs (IDATs, CELs) |
| Binary Alignment files (BAMs, de-multiplexed) – linked with participant de-identified ID | | Manifest file describing SNP or probe content on the genotyping array |
| Associated phenotypic data collected | Associated phenotypic data collected | Associated phenotypic data collected |
| Variant calling files (VCFs) | Final analyses BIOM files (at minimum must contain OTUs) | Final reports and analysis files generated |
| Mapping file indicating the relationship between the submitted files | Mapping file indicating the relationship between the submitted files | Mapping file indicating the relationship between the submitted files (completed Array Format template) |

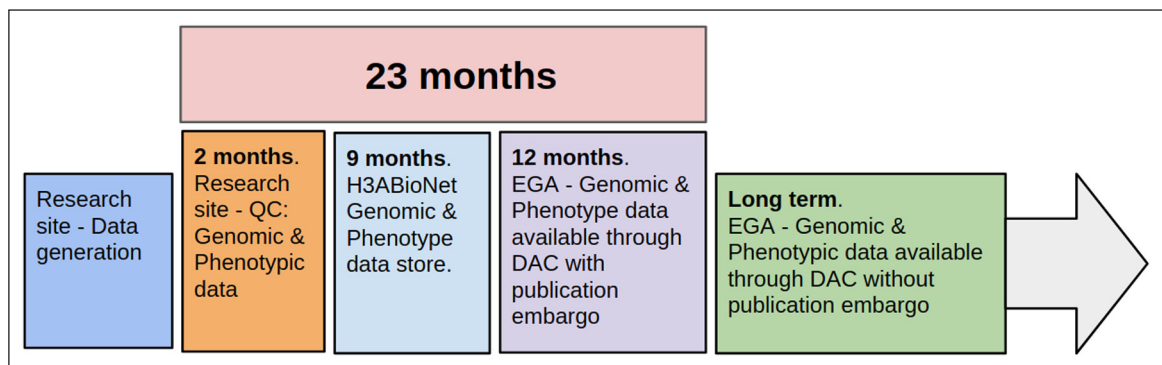


Figure 1: Timeline for submission of data to public repositories, extracted from the H3Africa Data Sharing, Access and release policy.

The data validation and transfer process from the data source to the H3Africa Archive can take some time –often from two to four months depending on the availability of storage space, the transfer mechanism used, the speed of internet connections and availability of technical human resources. From the date the data gets accepted into EGA, there is a 12 month publication embargo period through the DBAC. This means that the project which owns the data has 12 months to write and publish their papers with no threat of impingement on their work. Some journals require accession IDs before papers can be published, so the data needs to be in EGA prior to paper submission. When the 12 months are over, any researcher can access the data in EGA through the DBAC without a publication embargo.

2.4. Submission Engagement Process

H3Africa projects collect data from their sites or providers and store it at their hub for analysis. Engagement early on with the H3Africa projects and identification of the data managers and individuals who will be submitting data to the H3Africa Data Archive is beneficial in building up key stakeholder relationships. This engagement enables one to gauge how far the projects are with their timelines and provide an estimation of when data will be submitted to the H3Africa Data Archive enabling the adequate provisioning of resources. A remote meeting is arranged with the data submitters to determine what infrastructure and resources are available in terms of bandwidth, technical expertise and experience in using data transfer tools. The data submitters are then provided with a Data Submission pack and are encouraged to register their submission on the H3Africa Archive Dashboard in order to keep track of the various submissions and their current

status. The HDAT assists the data submitters in preparing and encrypting their data for submission through providing guidelines and a series of meetings (**Figure 2**).

2.5. Data Submission Request and Files

The information collected that is commonly referred to as the Data Submission Request (DSR) includes organization, abstract, dataset name, description, estimated deadline for submission, data type, institutional reference ethics code, phenotype variables, file types, size, number of samples, cases, controls, and link to GitHub code used to generate the data and analysis. Two additional files needed are the blank copy of the Case Report Form (CRF) or Questionnaire and a blank copy of the Consent Form used to collect the data. Projects also sign the H3Africa Archive Statement Agreement, this document gives H3ABioNet the right to validate and submit the data to the EGA.

The H3ABioNet Archive Team sends a submission pack to the data submitter after receiving the initial intent to submit data. The pack includes a copy of the DSR confirming the information from the project as well as the mapping files. These files vary depending on the kind of data, but should all include a phenotype data mapping file. The projects should all be collecting phenotype data, if they are not they specify this in the DSR. Phenotypes are mostly sex, ethnicity and country, but they are encouraged to include any other phenotype data collected in the Case Report Form (CRF).

3. Results

As part of its role in H3Africa, H3ABioNet agreed to host the H3Africa Archive to implement the consortium's Data Sharing, Access and Release policy. This required the development of data submission policies, guideline documents and data infrastructure that was both secure and scalable. Below we describe the results of this infrastructure development.

3.1. The H3Africa Data Archive Infrastructure

The H3Africa Data Archive physical infrastructure comprises three main components, a Landing Area server, a Vault server and Cold Storage. Initial scoping exercises were conducted to determine where the H3Africa Data Archive should be situated and to build a proof of concept. During the proof of concept (POC) stage, certain criteria were defined and online interviews conducted across H3ABioNet Consortium Nodes to assess their suitability to host the physical components of the H3Africa Data Archive. The Nodes are physically situated across various countries in Africa such as South Africa, Ghana, Nigeria, Morocco, Egypt, Senegal, Tunisia, Sudan, Kenya, Malawi, Uganda and Mauritius. These interviews focused on the following criteria:

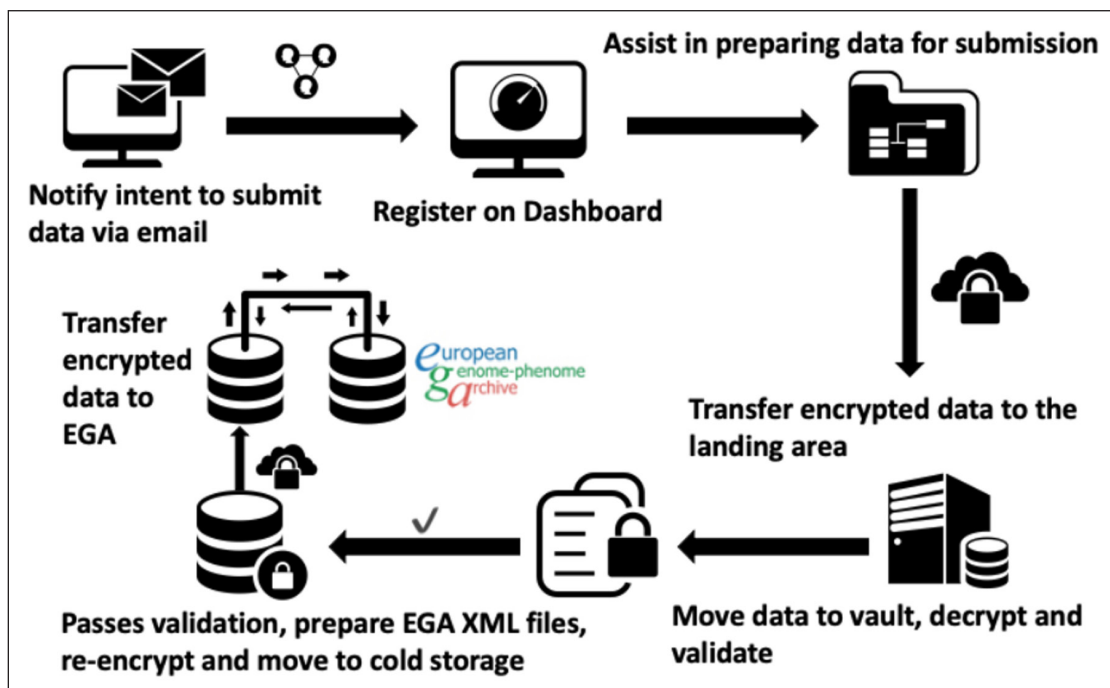


Figure 2: Diagram showing the process for submission of data to the Archive and EGA.

- Stable in country electrical supply
- Access to Uninterruptible Power Supplies (UPS) equipment
- Access to electrical generator hardware
- Existing IT technical human resources
- Existing IT infrastructure such as networking equipment, dedicated and secure datacenter room facilities
- Data backup infrastructures
- Ease of procurement

3.1.1. Landing Area

The Landing Area is where all the incoming and outgoing data is stored. All data on this server is always stored in encrypted format, with the public encryption key used for incoming data provided by the HDAT to the projects. The current storage capacity on the landing area is 50 Terabytes and is located in the institution's DMZ (demilitarized zone). The firewall rules in the DMZ are not as strict as those on internal firewalls. We encourage the use of GridFTP for transfer of large data sets to the archive (Ananthakrishnan et al. 2015). This protocol allows multiple connections to be open at once, masking TCP latency problems; in our experience this performs well in an African setting, and there are good, free services that provide this. Data transfer from the archive to the EGA uses a UDP-based service.

3.1.2. Vault

The Vault is a secure black-box server with tight access control policies in place. Data is validated in the Vault only, as it needs to be decrypted to do so. More details about validation are provided later. The HDAT works with the data submitter to fix any issues identified with the data during the validation phase. The server is also used for creating the XML schemas to submit the data to EGA. This is the only server where data is allowed to be decrypted. The Vault currently has access to 220 Terabytes of direct access storage (DAS).

3.1.3. Archival (Cold) Storage

The archival storage is where data is stored for the 9 month incubation period. A copy of the data encrypted with a separate H3ABioNet public/private key pair is kept in archival storage while a second copy is encrypted using the EGA public key and submitted to the EGA. All data in archival storage is replicated to off-site secure storage for redundancy purposes. The archival storage is expandable up to 500 Terabytes.

3.1.4. EGA Deposit Box

The EGA has made server space available for the H3Africa Consortium to deposit their data, also known as an EGA deposit box. Every entity submitted is given an accession ID, this includes samples, data sets, analyses, runs, experiments and others. Both the HDAT and EGA teams have access to the EGA data deposit box. For security purposes, all data submitted to the EGA data deposit box is encrypted using the EGA encryption key before being transferred.

3.2. Data Transfer

The transfer of data to the EGA can take quite some time, depending on the data size and Internet speed. **Table 2** shows a typical example of how long a data set takes to reach the EGA.

Data submitted via the Internet is ingested into the Vault from the Landing Area (**Table 2**, step 1). Likewise, data that is due to be submitted to the EGA is moved from the Vault to the Landing Area. Due to network design, the only method of moving data between the Vault and Landing Area is via the network. It is not

Table 2: Example speeds for time for moving data within and between the Archive and EGA.

| | From | To | Average Mbp/s (Megabits) | Average Mb/s (Megabytes) | Time to transfer (days) | Size |
|---|--------------------|---------------------------------|--------------------------------|--------------------------------|-------------------------------|--------|
| 1 | Vault | Landing Area | 120 | 15 | 6 | 8.9 TB |
| 2 | Other local server | EGA (Aspera) | 16 | 2 | 90 | 8.9 TB |
| 3 | Vault | Hard Drive (directly into port) | 200 | 25 | 3 | 8.9 TB |
| 4 | Landing Area | EGA (Aspera) | 240 | 30 | 2 | 8.9 TB |

currently possible to connect external USB storage to the Landing area. The slowest data transfer speeds are recorded when transferring data from a local server via the internal network to the EGA. This is largely due to the various firewalls and packet inspection tools implemented along the data transfer path. As expected, connecting a high speed USB storage device directly to the Vault server yields higher transfer rates. The data transfer rate shown in **Table 2** (step 3) was conducted using a USB 3.0 enabled storage device. The Landing Area is located in the institutional DMZ, as data transfers to and from this server have been optimised for Internet based data transfers which yields the fastest transfer speeds.

As is evident from **Table 2**, data transfers are a challenge when Internet speeds are slow and resources are limited. The data archive's preferred online data transfer mechanism is "Globus Online" which uses GridFTP (Foster et al. 2011), while EGA uses Aspera. The Aspera and Globus Online (GO) data transfer applications are optimised to efficiently and securely transfer data between two points on a public or private network (Madduri et al. 2014). Both applications have security and fault recovery built into the system which sets it apart from traditional data transfer methods such as FTP (file transfer protocol). The fault recovery measures work by setting checkpoints as data is successfully delivered to its destination. In the event of a network failure, when the transfer is restarted, GO or Aspera will pickup from the last checkpoint, compared to FTP which would require restarting the entire copy (Lappalainen et al. 2015).

3.3. Encryption

The primary aim of data encryption is to secure the genomic data (Mahdi et al. 2018). Encryption works by encoding data using a secret private key. The data is only accessible or readable when using the matching public key to decrypt it. Without the private key, the data is inaccessible, making it a suitable method of securing data whilst in transit across public networks such as the Internet. Encryption to EGA has a separate public key which is built into the EGACryptor tool. A major challenge of encryption is that it takes longer to encrypt or decrypt a file compared to using a standard password to protect the file. It also requires additional storage space up to three times the size of the raw data. This is not much of an issue when working with small data sets, but for larger data sets, such as genomic data, storage space for encryption becomes an important factor. To encrypt 1 Terabyte of data can take approximately 1 hour and 30 minutes. This varies depending on the file type and amount of resources used on the server at a particular time.

3.4. Archive Dashboard

The Archive Dashboard is a web application that was built to keep track of data submissions from the data submitters. The dashboard tracks all the progress of the submissions in an intuitive user friendly interface. A user is able to register, login and fill in a data submission request form. The HDAT will respond by assisting the data submitter with the submission pack and file formats. Funders or project managers can login to the dashboard with different access rights to view the progress of data from first engagement to submission to the EGA.

3.5. Current status of the Archive

At the time of writing, a total of 9 data sets have been submitted to the H3Africa Archive and the EGA. The total size of all the data sets submitted is 118.7 Terabytes with the average data set being 13.2 Terabytes. There are currently two studies that have been submitted to the EGA. The AWI-Gen Pilot Study (accession number: EGAS00001002482) is accessible via the EGA and the H3Africa Chip (accession ID: EGAS00001002976) is currently under embargo and expected to be accessible to the greater scientific community soon. More datasets are expected to be submitted to the H3Africa Data Archive in the near future.

4. Discussion and Conclusions

In order to implement the H3Africa data sharing policies, we developed what to our knowledge is the first human genomic data archive in Africa. There were many challenges encountered in the development of the infrastructure, most notably in data transfers when moving data around the globe and specifically across the African continent. Common challenges include:

- Researchers or data owners not wanting to share their data
- Communication issues
- Technical issues, such as slow internet speeds or expensive bandwidth
- Available server compute resources, available storage space or familiarity with data transfer technologies
- Data governance which restricts the movement of genomic data across borders

The H3Africa Archive, though developed to address internal needs of the consortium, provides a useful proof of concept for the possibility of establishing local EGA facilities. It was designed based on the EGA architecture, and ensures data security and conversion into EGA formats. A similar infrastructure could be used for other genomic data where an archive of files is required with built in secure storage, off-site replication and data transfer procedures. Our experience has demonstrated that significant long-term resources are required for such an infrastructure, including both human and computational. We also recognize the value in data sharing initiatives as researchers and funders move increasingly to an open science ethos. Researchers from the project sites can benefit tremendously from the data sharing capabilities of the Internet. In addition to having access to international data sets, by submitting their data to public data archives such as the EGA, they expose their research to the greater scientific community which in itself holds many benefits.

Competing Interests

The authors have no competing interests to declare.

References

- Ananthakrishnan, R, Chard, K, Foster, I and Tuecke, S.** 2015. Globus Platform-as-a-Service for Collaborative Science Applications. *Concurrency and computation: Practice & experience*, 27(2): 290–305. DOI: <https://doi.org/10.1002/cpe.3262>
- Aronson, SJ and Rehm, HL.** 2015. Building the foundation for genomics in precision medicine. *Nature*, 526(7573): 336–342. DOI: <https://doi.org/10.1038/nature15816>
- Christensen, KD, Dukhovny, D, Siebert, U and Green, RC.** 2015. Assessing the Costs and Cost-Effectiveness of Genomic Sequencing. *Journal of personalized medicine*, 5(4): 470–486. DOI: <https://doi.org/10.3390/jpm5040470>
- Cock, PJA, Fields, CJ, Goto, N, Heuer, ML and Rice, PM.** 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6): 1767–1771. DOI: <https://doi.org/10.1093/nar/gkp1137>
- de Vries, J, Tindana, P, Littler, K, Ramsay, M, Rotimi, C, Abayomi, A, Mulder, N and Mayosi, BM.** 2015. The H3Africa policy framework: Negotiating fairness in genomics. *Trends in Genetics*, 31(3): 117–119. DOI: <https://doi.org/10.1016/j.tig.2014.11.004>
- Dyke, SOM, Linden, M, Lappalainen, I, De Argila, JR, Carey, K, Lloyd, D, Spalding, JD, Cabili, MN, Kerry, G, Foreman, J, Cutts, T, Shabani, M, Rodriguez, LL, Haeussler, M, Walsh, B, Jiang, X, Wang, S, Perrett, D, Boughtwood, T, Matern, A and Flicek, P.** 2018. Registered access: Authorizing data access. *European Journal of Human Genetics*, 26(12): 1721–1731. DOI: <https://doi.org/10.1038/s41431-018-0219-y>
- Foster, I.** 2011. Globus online: Accelerating and democratizing science through cloud-based services. *IEEE Internet Comput*, 15: 70–73. DOI: <https://doi.org/10.1109/MIC.2011.64>
- Goldfeder, RL, Wall, DP, Houry, MJ, Ioannidis, JPA and Ashley, EA.** 2017. Human Genome Sequencing at the Population Scale: A Primer on High-Throughput DNA Sequencing and Analysis. *American Journal of Epidemiology*, 186(8): 1000–1009. DOI: <https://doi.org/10.1093/aje/kww224>
- H3Africa Consortium, Rotimi, C, Abayomi, A, Abimiku, A, Adabayeri, VM, Adebamowo, C, Adebisi, E, Ademola, AD, Adeyemo, A, Adu, D, Affolabi, D, Agongo, G, Ajayi, S, Akarolo-Anthony, S, Akinyemi, R, Akpalu, A, Alberts, M, Alonso Betancourt, O, Alzohairy, AM, Ameni, G, et al.** 2014. Research capacity. Enabling the genomic revolution in Africa. *Science*, 344(6190): 1346–1348.
- Hsi-Yang Fritz, M, Leinonen, R, Cochrane, G and Birney, E.** 2011. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Research*, 21(5): 734–740. DOI: <https://doi.org/10.1101/gr.114819.110>
- Lappalainen, I, Almeida-King, J, Kumanduri, V, Senf, A, Spalding, JD, Ur-Rehman, S, Saunders, G, Kandasamy, J, Caccamo, M, Leinonen, R, Vaughan, B, Laurent, T, Rowland, F, Marin-Garcia, P, Barker, J, Jokinen, P, Torres, AC, de Argila, JR, Llobet, OM, Medina, I and Flicek, P.** 2015. The European Genome-phenome Archive of human data consented for biomedical research. *Nature Genetics*, 47(7): 692–695. DOI: <https://doi.org/10.1038/ng.3312>
- Madduri, RK, Sulakhe, D, Lacinski, L, Liu, B, Rodriguez, A, Chard, K, Dave, UJ and Foster, IT.** 2014. Experiences Building Globus Genomics: A Next-Generation Sequencing Analysis Service using Galaxy, Globus, and Amazon Web Services. *Concurrency and computation: practice & experience*, 26(13): 2266–2279. DOI: <https://doi.org/10.1002/cpe.3274>

- Mahdi, MSR, Aziz, MMA, Alhadidi, D and Mohammed, N.** 2018. Secure similar patients query on encrypted genomic data. *IEEE journal of biomedical and health informatics*. DOI: <https://doi.org/10.1109/JBHI.2018.2881086>
- Mulder, N, Abimiku, A, Adebamowo, SN, de Vries, J, Matimba, A, Olowoyo, P, Ramsay, M, Skelton, M and Stein, DJ.** 2018. H3Africa: Current perspectives. *Pharmacogenomics and personalized medicine*, 11: 59–66. DOI: <https://doi.org/10.2147/PGPM.S141546>
- Mulder, NJ, Adebisi, E, Adebisi, M, Adeyemi, S, Ahmed, A, Ahmed, R, Akanle, B, Alibi, M, Armstrong, DL, Aron, S, Ashano, E, Baichoo, S, Benkahla, A, Brown, DK, Chimusa, ER, Fadlelmola, FM, Falola, D, Fatumo, S, Ghedira, K, Ghouila, A and H3ABioNet Consortium, as members of the H3Africa Consortium.** 2017. Development of bioinformatics infrastructure for genomics research. *Global heart*, 12(2): 91–98. DOI: <https://doi.org/10.1016/j.gheart.2017.01.005>
- Mulder, NJ, Adebisi, E, Alami, R, Benkahla, A, Brandful, J, Doumbia, S, Everett, D, Fadlelmola, FM, Gaboun, F, Gaseitsiwe, S, Ghazal, H, Hazelhurst, S, Hide, W, Ibrahimi, A, Jaufeerally Fakim, Y, Jongeneel, CV, Joubert, F, Kassim, S, Kayondo, J, Kumuthini, J and H3ABioNet Consortium.** 2016. H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa. *Genome Research*, 26(2): 271–277. DOI: <https://doi.org/10.1101/gr.196295.115>
- Popejoy, AB and Fullerton, SM.** 2016. Genomics is failing on diversity. *Nature*, 538(7624): 161–164. DOI: <https://doi.org/10.1038/538161a>
- Prokop, JW, May, T, Strong, K, Bilinovich, SM, Bupp, C, Rajasekaran, S, Worthey, EA and Lazar, J.** 2018. Genome sequencing in the clinic: The past, present, and future of genomic medicine. *Physiological Genomics*, 50(8): 563–579. DOI: <https://doi.org/10.1152/physiolgenomics.00046.2018>
- Shabani, M, Dyke, SOM, Joly, Y and Borry, P.** 2015. Controlled Access under Review: Improving the Governance of Genomic Data Access. *PLoS Biology*, 13(12): e1002339. DOI: <https://doi.org/10.1371/journal.pbio.1002339>


How to cite this article: Parker, Z, Maslamoney, S, Meintjes, A, Botha, G, Panji, S, Hazelhurst, S and Mulder, N. 2019. Building Infrastructure for African Human Genomic Data Management. *Data Science Journal*, 18: 47, pp. 1–8. DOI: <https://doi.org/10.5334/dsj-2019-047>

Submitted: 31 January 2019

Accepted: 12 September 2019

Published: 26 September 2019

Copyright: © 2019 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiq Press.

OPEN ACCESS 