



Genomic Research Data Generation, Analysis and Sharing – Challenges in the African Setting

Nicola Mulder, Clement Adebamowo, Sally Adebamowo, Oladimeji Adebayo, Osimhiarherhuo Adeleye, Mohamed Alibi, Shakuntala Baichoo, Alia Benkahla, Faisal Fadlelmola, Hassan Ghazal, et al.

► To cite this version:

Nicola Mulder, Clement Adebamowo, Sally Adebamowo, Oladimeji Adebayo, Osimhiarherhuo Adeleye, et al.. Genomic Research Data Generation, Analysis and Sharing – Challenges in the African Setting. CODATA Data Science Journal, Committee on Data for Science and Technology (CODATA), 2017, 16 (49), pp.1 - 15. 10.5334/dsj-2017-049 . pasteur-02010733

HAL Id: pasteur-02010733

<https://hal-riip.archives-ouvertes.fr/pasteur-02010733>

Submitted on 7 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



RESEARCH PAPER

Genomic Research Data Generation, Analysis and Sharing – Challenges in the African Setting

Nicola Mulder¹, Clement A. Adebamowo², Sally N. Adebamowo², Oladimeji Adebayo³, Osimhiarherhuo Adeleye⁴, Mohamed Alibi⁵, Shakuntala Baichoo⁶, Alia Benkahla⁵, Faisal M. Fadlelmola^{7,8}, Hassan Ghazal⁹, Kais Ghedira⁵, Alice Matimba¹⁰, Ahmed Moussa¹¹, Zahra Mungloo-Dilmohamud⁶, Mayowa O. Owolabi³, Fouzia Radouani¹², Charles N. Rotimi¹³, Dan J. Stein¹⁴, Oussama Souiai⁵ and as members of the H3Africa Consortium

¹ Computational Biology Division, Department of Integrative Biomedical Sciences, Institute for Infectious Disease and Molecular Medicine, University of Cape Town, ZA

² Department of Epidemiology and Public Health, University of Maryland Greenebaum Comprehensive Cancer Center, University of Maryland School of Medicine, Baltimore, MD, 21201, US

³ Department of Medicine, University College Hospital, Ibadan, NG

⁴ Federal Medical Centre, Abeokuta, NG

⁵ Université Tunis El Manar, Institut Pasteur de Tunis, Laboratory of Bioinformatics, BioMathematics and BioStatistics, TN

⁶ University of Mauritius, MU

⁷ Centre for Bioinformatics and Systems Biology, Faculty of Science, University of Khartoum, SD

⁸ Future University of Sudan, SD

⁹ Laboratory of Physiology and Genetics, University Mohamed Premier, MA

¹⁰ Department of Clinical Pharmacology, University of Zimbabwe, Harare, ZW

¹¹ SDET, ENSAT, Abdelmalek Essaadi University, MA

¹² Institut Pasteur du Maroc Casablanca Morocco, MA

¹³ Center for Research on Genomics and Global Health, National Human Genome Research Institute, NIH, Bethesda, MD 20892, US

¹⁴ Department of Psychiatry, University of Cape Town and MRC Unit on Risk and Resilience in Mental Disorders, ZA

Corresponding author: Nicola Mulder (nicola.mulder@uct.ac.za)

Genomics is the study of the genetic material that constitutes the genomes of organisms. This genetic material can be sequenced and it provides a powerful tool for the study of human, plant and animal evolutionary history and diseases. Genomics research is becoming increasingly commonplace due to significant advances in and reducing costs of technologies such as sequencing. This has led to new challenges including increasing cost and complexity of data. There is, therefore, an increasing need for computing infrastructure and skills to manage, store, analyze and interpret the data. In addition, there is a significant cost associated with recruitment of participants and collection and processing of biological samples, particularly for large human genetics studies on specific diseases. As a result, researchers are often reluctant to share the data due to the effort and associated cost. In Africa, where researchers are most commonly at the study recruitment, determination of phenotypes and collection of biological samples end of the genomic research spectrum, rather than the generation of genomic data, data sharing without adequate safeguards for the interests of the primary data generators is a concern. There are substantial ethical considerations in the sharing of human genomics data. The broad consent for data sharing preferred by genomics researchers and funders does not necessarily align with the expectations of researchers, research participants, legal authorities and bioethicists. In Africa, this is complicated by concerns about comprehension of genomics research studies, quality of research ethics reviews and understanding of the implications of broad consent, secondary analyses of shared data, return of results and incidental findings. Additional challenges with genomics research in Africa include the inability to transfer, store,

process and analyze large-scale genomics data on the continent, because this requires highly specialized skills and expensive computing infrastructure which are often unavailable. Recently initiatives such as H3Africa and H3ABioNet which aim to build capacity for large-scale genomics projects in Africa have emerged. Here we describe such initiatives, including the challenges faced in the generation, analysis and sharing of genomic data and how these challenges are being overcome.

Keywords: Genetics; genomics; bioinformatics; Africa; health

Introduction

Broadly speaking, genomics is the study of the DNA that makes up the genomes of organisms, including sequencing and analysis of the structure and function of these molecules. Genomics is a powerful tool for the study of human, pathogen, plant and animal evolutionary history, pharmacogenomics and diseases through the analysis of genetic variations. Genomic data generation has achieved significant economy of scale in the context of genotyping and sequencing. This success story of decreasing costs of genomic technologies and the resulting increased data size and complexity is posing new challenges to scientists. The infrastructure and skills to manage, store, analyze and interpret genomic data are not keeping pace with the ever increasing data generation capabilities. Phenotypic characterization and collection of biological samples require enormous effort and input from multiple stakeholders. As a result, once data is generated, researchers are reluctant to immediately share it, because they want the opportunity to exploit the data given their efforts. The concern about sharing research data is particularly strong among African scientists who often lack the capacity to generate and analyze the genomics data arising from their samples. The outcome is dissatisfaction about the balance of recognition and benefits between primary data collectors and, primary and secondary genomics analysts in authorship, patents and other outcomes of research projects. For example, some of the major publications on genome sequences from African individuals have been led from abroad (Tishkoff et al. 2009; Lambert and Tishkoff, 2009; Schuster et al. 2010).

Given the sensitive nature of human genetic data and the need to share samples and data across countries in large collaborative research projects, the nature and scope of informed consent raise several ethical, legal and social concerns. This is particularly so in the African context where novel and unique circumstances and opportunities may arise from the processes of informing and receiving consent for genomic research. The consent preferred by genomics researchers and funders, including the stipulation to share data may be problematic for other researchers, research participants, legal authorities and bioethicists. Comprehension of genomics research projects by participants, understanding the implications of broad consent and secondary analyses of shared data and samples, quality of research ethics reviews, return of genomics research results in the face of uncertainty about the meaning and lack of resources for verification, and incidental genomics findings pose additional ethical challenges for genomics research in Africa. For non-human genomic data, the ethical challenges are less, but there are potential commercial opportunities to be gained for example in the design of new drugs and vaccines against pathogens, which also form a barrier to open data sharing.

The infrastructural challenges posed by the increasing data generation capabilities of genomic technologies is particularly acute for researchers across the African continent. The highly specialized skills, extensive and expensive computing infrastructure, broadband internet access, secure cloud computing and uninterrupted power supply are not readily available across the continent. New initiatives are being established which aim to enable large-scale genomics projects in Africa to study the genetic basis of human history and diseases, but also have strong capacity building elements. One such effort is the Human Heredity and Health in Africa initiative (H3Africa: www.h3africa.org) designed to generate large and complex genomics datasets from multiple ethnic groups across Africa (H3Africa consortium 2014). At the core of the informatics activities of H3Africa is the H3ABioNet (www.h3abionet.org) (Mulder et al. 2015). This Pan African bioinformatics network is building bioinformatics capacity to enable genomics data analysis on the continent. Here, as members of H3Africa projects and the H3ABioNet network, we describe the challenges associated with data generation, analysis and sharing in Africa, and how these initiatives and networks are working towards overcoming these challenges.

Genomics data generation

Genomics, a Big Data science, is generating ever increasing amounts of data, mostly from sequencing thousands of human and non-human genomes. In the mid 2000's several new sequencing technologies based on massive parallel DNA sequencing approaches, referred to as next generation sequencing (NGS) emerged, which allow genomic scientists to sequence faster and cheaper. This has revolutionized our ability to interrogate the genomes of thousands of humans and other organisms, facilitating novel insights into biology, medicine and evolutionary history (Goodwin, McPherson and McCombie 2016). These efforts require efficient data acquisition, storage, distribution, and analysis. Storage, distribution and analysis of data are core activities associated with any large amount of data, but data acquisition in genomics is highly distributed and involves heterogeneous formats. There is a need to systematically organize this data and to disseminate the resulting information through technically sound means to provide opportunities to academics and researchers globally. This information dissemination can facilitate advances in biomedical research for the improvement of health (Stephens et al. 2015). Much of the data generated by NGS applications and other omics (including genomics, transcriptomics, proteomics and other high-throughput methods) technologies are housed in public databases (Rung and Brazma 2013). These databases are either general (e.g. Sequence Read Archive, Array express, Gene Expression Omnibus) or dedicated to one research area (e.g. OncoPrint for cancer research, plexDB for plant research). They are openly accessible for use by any researcher, though some, notably those containing human data, require access approval from specialist committees.

Genomic technologies are still relatively unaffordable in Africa. Generation of genomic datasets for human health and populations research has been limited, with South Africa, Ghana and Kenya being the most studied (Adedokun et al. 2016; HuGE database). However, more recently, various African institutions are equipping themselves with next generation sequencing platforms and setting up systems for genomic data generation. H3Africa is the largest project to date with a focus on generating large scale genomic sequence data by scientists in Africa for human health research. There are different data types being generated by H3Africa projects, including genotyping by arrays, whole human genome or exome sequencing, targeted sequencing of human or pathogen genes, and 16S rRNA sequencing for microbiome analysis. The capacity for its storage and analysis is the subject in question and brings about the urgent need for improving resources and capacity for data sharing and analysis in Africa.

Data transfer and storage

Until recently, African institutions were not generating large genomics datasets alone, due to high cost and unavailability of equipment. Most were not involved in genomics research at all and the few that were generally outsourced their sequencing work or downloaded data from publicly available datasets. In both cases, a new challenge arose with transfer of large datasets in environments with limited infrastructure. Big data transfer is an essential service particularly when collaborating across multiple sites. For many African countries, data transfer is expensive, unreliable, insecure, and difficult to monitor. For example, it took 5 months to transfer 140TB of sequence data from the USA to South Africa using currently available transfer resources. With reliable internet, sufficient storage in place and constant monitoring, this should have taken 2 months, but delays were caused by low bandwidth and internet down times. Using File Transfer Africa (<http://filetransferafrica.co.za/3-5tb-and-2-6-million-files-uploaded-to-europe/>), 3.5 terabytes were transferred at approximately 1 gigabyte per minute which equates to approximately, 60 gigabytes per hour or 1 terabyte every 12 hours. This meant the file transfer took a combined time of 40 hours 'on the wire'. Based on this, if we use File transfer Africa, the transfer of 140TB should take approximately 66.7 days, or 2 months. However, many African Institutions do not have reliable internet connections and in some cases power supply is erratic because of weather-related damage to municipal infrastructure, and low levels of investment in public services such as electricity supply. There is also substantial lack of computer infrastructure and IT personnel to handle the large amount of data.

H3ABioNet has invested a lot of effort in infrastructure and skills development to enable big data transfer in Africa. There were challenges in trying to identify a single data transfer solution that can be used regardless of the operating system and the networking status of the institution. H3ABioNet started the NetMap project to map the internet speeds between its nodes (member institutions located throughout the continent), and implemented a unified transfer solution using Globus Online (Foster 2011). Technical staff at nodes were trained to maintain their computing infrastructure and support their researchers with data transfer needs. How-To guides were developed to help technical staff and ensure the sustainability of the infrastructure.

Genomic data storage has also been a challenge. This includes the primary data, temporary files created during analysis, and the final dataset generated from research projects. Storing this large amount of data in federated but connected resources is a challenge in its own right. Many African and other Low and Middle Income Country (LMIC) institutions lack a well maintained and organized storage unit with built in redundancy that can ensure the security of research data. Apart from technical challenges, data organization and tracking is also a challenge. Genomic data can vary between several huge files and thousands of small files. Organizing them in ways that make them easily accessible and identifiable to researchers can be difficult and requires careful curation of the metadata.

There is enormous potential for scientific discovery when datasets from multiple projects are merged, particularly with genomic research that requires large sample sizes. In order to facilitate this, datasets need to be harmonized to ensure the metadata is consistent (same terms mean the same things), and data within files are easily searchable and findable. Within the H3Africa consortium alone, the clinical data has proven to be quite heterogeneous. Even when similar clinical observations have been captured, they are often not comparable due to differences in phrasing of questions or measurements. Therefore special effort is being invested in harmonizing data by mapping them to existing biomedical ontologies, which is essential for efficient data sharing. For example, curators have assessed case report forms for H3Africa projects to determine compatibility of data collected with similar but non-identical questions. The data is then being mapped to the phenotype, disease and experimental factor ontologies. Additionally, a recommended minimal set of questions based largely on PhenX measures, (<https://www.phenxtoolkit.org/>) have been drafted for new projects starting recruitment of participants. More effort should be invested in future projects to ensure that better data harmonization strategies and rules are adopted and enforced.

Physical and infrastructural challenges of large scale genomic studies in Africa

There are few high performance computing centers in Africa outside of South Africa, although this is slowly changing. Additionally, the use of Cloud computing has not been practical for large datasets located on the continent due to slow internet speeds. Data security is also an issue for human data. Another challenge is the heterogeneous nature of the multitude of bioinformatics tools available. Many bioinformatics algorithms and software have been written by students who move on to other projects, so they remain unmaintained and unsupported. The field moves so rapidly that tools and pipelines constantly need to be updated. One advantage though is that most bioinformatics developers keep their code open source and available in *github* for the community to use and develop further. Nevertheless, using these tools for big data analysis in biomedical research requires familiarity with computer programming for data manipulation and bug fixing, biostatistics and use of analysis software. There is currently limited availability of skills to process and interpret big data in Africa. In anticipation of the development of an African Research Cloud, H3ABioNet has attempted to ease the barrier to accessing tools and computing by developing workflows for common data types that can be deployed on local high performance computing clusters or clouds. The network is also using courses and hackathons to build skills in the development and running of these workflows.

Sharing of genomic data

An ongoing challenge in the context of genomics research, is the relative novelty of data sharing practices in Africa, partly because African scientists have not been large generators of genomic data in the past. This novelty frequently translates into considerable concerns by ethics committees to approve genomics research protocols (Ramsay et al. 2014; De Vries et al. 2015a; De Vries et al. 2016) – although admittedly this is something that is more pertinent to the sharing of genomic samples than it seems to be the case for genomic data. Data sharing with a wide range of users is increasingly required by funding agencies, to maximize the use of data generated from individual research institutions or consortia. Sharing genomics data provides an opportunity to verify original analyses, improve reproducibility, test new hypotheses and combine datasets from different sources to achieve higher statistical power. Regardless of these potential benefits, sharing genomic data has provoked concerns for research participants and researchers. These concerns include endangering the privacy of the data subjects (Gymrek et al. 2013; Homer et al. 2008) and downstream uses of data (Shabani et al. 2016), which were not addressed by the initial informed consent process.

To address some of these concerns, several resources have been put in place, including setting up Data Access Committees (DACs), which are responsible for data release to external requestors based on legal, ethical and scientific eligibility. In a survey of DAC members involved in reviewing access requests for genomic

data available in databases of Genotypes and Phenotypes (dbGaP) and the European Genome-phenome Archive (EGA) on their experiences and attitudes to the tools and mechanisms for access review and its adequacy in fulfilling the goals of controlled-access model of data sharing, the researchers concluded that DAC members and experts were ambivalent about the effectiveness and consistency of the review procedures and oversight process (Shabani et al. 2016) Therefore, data sharing policies need to be structured in ways that address the ethical and access control concerns of all research stakeholders. H3Africa has developed a data sharing, access and release policy and established a Data and Biospecimen Access Committee (DBAC) with guidelines on its composition and role (H3Africa guidelines and policy documents are available at: <http://www.h3africa.org/consortium/documents>). The data access policy stipulates that genomic and phenotype data must be submitted to public repositories in a timely manner. These timelines are discussed later.

Reluctance to share data can also be a major obstacle for academic or commercial reasons. Genomic data from the health or agricultural sectors may have commercial potential if, for example, it can lead to the development of novel vaccines, therapies or pesticides (genome analysis can identify potential novel targets for drugs or vaccines). While open access to data accelerates science, ownership of intellectual property and the benefits derivable from that may be barriers to sharing. African researchers are aware of the financial rewards and recognitions that have accrued to researchers in other parts of the world from the outcome of their research and are concerned about not receiving similar benefits from their research or credit for sharing their hard-won data (van Panhuis et al. 2014). African researchers are particularly concerned about being perceived and treated as mere data collectors who do not make sufficient intellectual contributions worthy of similar levels of recognition and benefits received by other members of the research consortia. Many African institutions lack personnel and technical resources that can match those in High Income Countries (HIC) and enable them to quickly mine their own data and generate publications, patents and other benefits before the data become publicly available. Considerations therefore need to be given to extended periods of protected access for African institutions that is cognizant of this inequity in resources when timelines for data sharing are being determined.

Challenges to re-use of publicly available data

Even though several pan-African consortia, such as H3Africa, have aimed to strengthen the ability of African institutions to generate their own experimental data, these efforts remain limited by available funding. The concept of open data or open science that can be shared, freely used and reused by anyone for any purpose (The Open Knowledge Foundation: <http://opendefinition.org/>) provides a useful alternative that helps bioinformaticians and other scientists to overcome the lack of access to their own experimental data. Public data reuse facilitates addressing of new biological questions (Poldrack, and Gorgolewski 2014) and exploring secondary hypotheses that were not investigated in the original studies (Chokshi et al. 2006; Goodman 2015; Van Horn and Ball 2008); developing and evaluating new methods; producing new products and services enabling inter- and transdisciplinary research (Tenopir et al. 2011; Gaheen et al. 2013); implementing meta-analysis of data with the flexibility to include data and samples from different platforms (Rung and Brazma 2012) for making new observations that could not be detected using individual data sets; and integrating and analyzing several primary datasets in order to acquire new knowledge and/or build a new data resource (Rung and Brazma 2012). However, the reuse of data requires availability of sufficient information on the data generation process and experimental design to ensure its use in a scientifically appropriate way, as well as careful interpretation of the results.

Data repositories provide infrastructural solutions that enable scientists to share and reuse public data. The Wellcome Trust and the National Institutes for Health (NIH) have made large investments in sustainable infrastructures for genomic data (Pisani and AbouZahr 2010). Several online repositories and archives offer the possibility to store, access, use and reuse research and scientific inputs and outputs. Such platforms speed up the transfer of knowledge among researchers and across scientific fields, and open up new ways of collaborating that can produce new knowledge, products and services. In addition, a growing number of funding agencies and publishers around the world are advocating and enforcing data management for open access. Despite the efforts made by data providers to make 'omics' data obtainable by everyone and simple to access, there are still several challenges facing African researchers to properly exploit the data. Technically, there is the issue of having a sustainable transfer channel to get the data on site, and there is generally a lack of skills and African infrastructural support for sharing, storing, managing, archiving and retrieving data (Robinson 2014; Miller 2015). Also, shared data with an inappropriate data presentation format are less useful (Ferguson et al. 2014). Most scientists would like to access data from others, but are rarely willing

to disseminate their own data, except where intellectual property is recognized and rewarded. In addition, few grant-funded investigators want to spend precious time and funds preparing data for someone else's research, even in exchange for authorship credit. There are a number of ethical and legal issues: data from some older studies may not conform to widespread reuse as they may not have had consent forms that allow free sharing of data with other researchers. Furthermore, even for new data with consent, anonymization of patients is not guaranteed, leaving researchers vulnerable to failure to preserve patients' privacy.

Ethical challenges in Sharing of Genomic Data

Whilst there are a number of key ethical challenges relating to the sharing of genomic data, one of the most pertinent of these relates to how to promote fairness and equity in sharing. Whilst several international policies advocate for release of data immediately after curation, other policies recognize that this may disadvantage investigators based in LMICs who do not (yet) have the resources or capacity to analyze data rapidly (Bull 2016, Bull et al. 2015a). The concern is that rapid sharing could lead to unfair collaboration practices and promote experiences of exploitation. Indeed, past collaborations may have offered little opportunity for African scientists to intellectually engage in or lead African health research. Adedokun and colleagues (Adedokun et al. 2016) performed an analysis on 508 articles published between January 2004 and December 2013 using data from Sub-Saharan Africa (SSA) to assess the contributions of SSA scientists to genomics research involving African participants. While the majority of the publications (91.1%) had at least one author affiliated with an African institution, 8.9% did not include any author affiliated with an African institution. Less than the half (46.9%) of the publications had a first author from an African institution while the remaining proportion had a foreign scientist as first author (Adedokun et al. 2016). Data sharing that does not meaningfully engage African researchers, risks being ill-suited to the needs of African populations and may even lead to inventions that are not relevant to Africans (Ramsay et al. 2011). Another concern is that researchers who are not based in Africa may not be able to meaningfully interpret research findings which may aggravate existing stigma (Bull et al. 2015b, De Vries et al. 2014).

In order to address this problem, H3Africa adopted three mechanisms to ensure that H3Africa researchers have at least a fair chance to meaningfully use genomic resources (De Vries et al. 2015b, De Vries et al. 2015c). The first is to only make resources available after nine months to allow investigators time to analyze data and submit manuscripts for publication. A second mechanism involves adding an additional twelve month embargo period for shared data so that H3Africa investigators can reserve research questions that they will address using data that they have generated within a reasonable timeframe while other investigators refrain from publishing on those topics. A third mechanism is to outline specific requirements for secondary use of samples, and in some cases data. For instance, some data can only be used for studies on specific diseases, such as for cardiovascular research. For samples, secondary use must involve meaningful capacity building and involvement of African researchers for the first two years of access. New data generated off the samples are required to be submitted to the EGA under the H3Africa project.

Protection of participants

Protection of participants is very important in genomic research and data sharing. Informed consent is an important mechanism to avoid potential exploitation of research participants and protect their rights and well-being. Valid consent is a process rather than a simple one-off matter of signing a form, and participants in genomics research may need to be engaged multiple times during the research project to ensure their continued voluntary consent to all parts of the research. The consent process needs to be designed in a way that is culturally appropriate and understandable (Lemmens 2015). Investigators examining consent to genomic research in African settings have identified a number of challenges in communicating study goals, methods and procedures. (Traore 2015), These include linguistic and conceptual barriers to comprehension of the research process, voluntariness, relationship between research and clinical practice, broad consent and the potential consequences of future unspecified research using samples and data (Chokshi 2007; McGuire and Beskow 2010). Empirical evidence from across the continent shows that the majority of research participants do not grasp these concepts of genomics or are unable to explain what they mean (Tindana and De Vries 2016), although there is notable variation with regards to the location (urban/rural) and demographics of research participants (Marshall et al. 2014). There is often difficulty in finding local equivalent words for genomics terms and explaining these novel, unfamiliar and highly technical subjects in local languages (Munung et al. 2016; Traore et al. 2015; Tindana et al. 2012). The process of working out comprehensible language is crucial, whether in writing or verbally. Chokshi et al (2007) suggest that

this process should involve researchers, institutional review bodies (IRB), funders and communities jointly determining commonly accepted language, oral and written, for particular concepts. In addition to its classical elements, it is recommended that informed consent for genomic research should address four major elements associated with explanations of data and sample sharing: i) authorities (e.g. DACs and IRB/RECs) deciding on reuse of samples, (ii) restrictions on secondary use (e.g. when providing conditions for broad consent) (iii) reasons for storage and sharing of data and samples and (iv) the role of biobanks, which may include timeline for sample storage (Munung et al. 2016).

Community engagement (CE) plays an important role in extending the ethical principle of respect for persons to entire communities, avoiding exploitation, and building trust between researchers and the communities involved in research (Lakes et al. 2014; Tindana et al. 2011). CE in the context of genomic research provides opportunities for informing and educating communities about genomics and genomic research, and exchanging information between the research team and potential research participants about the research process over a period of time (Tindana et al. 2015). Drawing an example from ongoing CE and outreach activities within the phenomics core of the Stroke Investigative Research and Education Network, SIREN (a member of the H3Africa Initiative), the CE framework design includes i) development and implementation of a Community Advisory Board (CAB) within each site to guide the ongoing research activities and research dissemination within communities, and ii) public outreach to communities and engagement with a focus on explaining the study, its objectives, expectations and to invite active participation. This community-based participatory research, in addition to other advantages, has allowed for the development of trusting community-researcher relationships and guided the researcher and team in disseminating findings and translating research into practice and policy (Jenkins et al. 2016). In some research settings, CE may be conducted prior to data collection, while in others it may continue throughout the duration of a study. CE has been shown to enhance understanding of research goals and procedures particularly with the complexities involved in genomic studies. It also provides an avenue for feeding back research findings to participants and communities (Rotimi et al. 2007; De Vries et al. 2011). Engaging the community in the course of the research grants members of the community an input into the research through their leaders. Therefore, ideally, CE activities should occur prior to, during and after a research project.

Guidelines and requirements of different ethics boards

International best practices, law and statutes of many African countries require independent ethics committees (institutional review boards (IRBs) or research ethics committees (RECs)), to provide third party review of research activities involving human subjects (Kass et al. 2007a). Although most African countries have some form of research review process, some still do not have national RECs (Nyika et al. 2009; Kirigia et al. 2005). Furthermore, IRB/RECs requirements and guidelines vary across African borders with local challenges (McWilliams et al. 2003). Recently, De Vries and collaborators (De Vries et al. 2017) conducted a comprehensive analysis on 30 existing ethics guidelines, policies and other similar sources from 22 African countries involved in H3Africa in order to better understand the ethics regulatory landscape around genomic research and biobanking in Africa. It was shown that the type of ethics guidance sources varied tremendously across African countries, including standard operating procedures (SOPs); national guidelines for health research; national guidelines for genetic research; ministerial decrees and laws. Among the countries included, only Malawi, Nigeria and South Africa had specific national or local guidelines for genomic and/or biobanking research. While the informed consent topic is discussed in all reviewed guidelines, the way in which it is described differs from being more abstract, such as the case of Kenya's guidelines, to very detailed, as is the case of Malawi's guidelines. The use of broad consent for future unspecified uses seems to be prohibited only in Zambia, Malawi and Tanzania, while the majority of countries (Benin, Ghana, Guinea, Kenya, Lesotho, Mauritius, Namibia, Swaziland, Togo and Zimbabwe) neither prevent nor promote broad consent. The guidelines of the remaining countries (Botswana, Sierra Leone, Senegal, Uganda, Cameroon, South Africa, Sudan, Rwanda, Nigeria and Ethiopia) allow consent for future unspecified research, but with conditions attached. The storage of samples is allowed in all the countries studied, however few of them offer specific guidance on the timeframe for storage, such as Zambia, in which samples can only be stored for a period of 10 years and permission is needed for a longer period. In Malawi, samples can only be stored for five years and in Zimbabwe, extraterritorial storage of samples beyond the study period is prohibited.

In contrast to sample storage, the export of samples is tightly controlled in many African countries. Indeed, guidelines from twelve countries (Ethiopia, Lesotho, Nigeria, Rwanda, Botswana, Malawi, South Africa, Zambia, Cameroon, Kenya, Uganda and Zimbabwe) address the export of samples and require approval from

one or more national agencies. Concerning sample re-use, only 14 countries (Botswana, Ghana, Ethiopia, Rwanda, Uganda, Kenya, Nigeria, Senegal, Sudan and Tanzania) out of 22, require approval from an ethics committee. The other countries are silent on whether approval from an ethics committee is required. Nine countries (Botswana, Cameroon, Ethiopia, Ghana, Kenya, Rwanda, Uganda, Zambia and Zimbabwe) specifically endorse international collaboration. Botswana, Kenya and Ugandan guidelines stipulate that export of samples is only allowed when there is no capacity to conduct the same analyses in the home country while, in Tanzania, if the local technology for the analysis exists, the researcher must explain why the samples need to be sent out of the country. Finally with respect to data sharing, only guidelines from Cameroon, Tanzania and Ethiopia mention data sharing and require a data sharing agreement to be submitted as part of the ethics application. Cameroon and Tanzania require review of all secondary studies by the ethics committee (De Vries et al. 2017).

General research ethics challenges in Africa include concerns about independence of RECs, inadequate funding and inadequate qualified staff in a background of weak health system infrastructures (Kass et al. 2007a; Nyika et al. 2009; Coleman and Bouësseau 2008). There is widespread poor capacity to handle ethical issues with serious implications for efficient functioning of RECs (Marzouk et al. 2014; Shabrawy et al. 2014). This poor capacity among RECs on the continent casts some doubt on their capacity to act as partners in genomics research in some places, in addition to the grave implications of making poorly informed decisions on applications (Tindana et al. 2012; Wright et al. 2013). There are also issues of poor representation in committees, poor understanding of the role of RECs, an inadequate number of properly constituted RECs, low standard of the RECs, and inadequate compensation for time on the committees for members (Coleman & Bouësseau 2008; Kirigia et al. 2005; Kass et al. 2007b; Ikingura et al. 2007). Therefore there is an urgent need for strengthening capacity for research ethics committees and for ethics research which could aid in development of national and regional policies which can support data sharing in Africa.

Although sharing of biospecimens is well established and enabled through material transfer agreements (MTAs), guidelines and mechanisms for sharing of genomics data is still evolving. Data protection laws differ across countries (<https://www.dlapiperdataprotection.com>) and could pose challenges if gaps identified in these laws are not recognized and addressed. The laws need to provide guidelines and policies that ensure success of consortium and collaborative research projects while ensuring that genomics data sharing is implemented in ways that are compatible with national laws and interests. Studies among H3Africa researchers, such as that from De Vries et al (2017) mentioned above, are underway to understand these gaps and improve alignment with national laws and guidelines. Unsurprisingly, there is no uniform informed consent template across different countries for genomic studies because there are wide cultural and ethnic disparities that make broad consent for data sharing difficult to apply across multinational studies (Wright et al. 2014). There are legal and ethical challenges to the transition from narrow consent, which is the usual practice, to broad consent, which is necessary in emerging fields, such as genomics research (Munung et al. 2016; Tindana et al. 2014; De Vries et al. 2015c). There are additional layers of bureaucratic obstacles in some countries, due to the strict requirements for Materials Transfer Agreements (MTAs). These are in place to protect the loss and exploitation of national samples, but can have an impact on the ability to send genetic materials outside the country (Federal Ministry of Health 2007; Staunton & Moodley 2013). Since data generation for many genetic studies is likely to be done outside the African shores, MTAs can thus impact data generation and transfer. In some cases MTAs are not clear about ownership of data or samples and how these would be governed especially where funding belongs to an external institution. Anecdotally, it may be stated that all materials belong to the government of the country where samples are obtained. However the international collaborator who is often well-resourced may dispute this ownership citing the significant funding contributions. Therefore there is need for development of guidelines on how these issues may be addressed. The Ethics and Governance Framework for Best Practice in Genomic Research and Biobanking in Africa recommends that MTAs should outline directions for handling commercializable outputs including benefit sharing arrangements.

Skills development for Big Data

Sustainable human resources

It is important to emphasize the challenge of sustaining ongoing big data analysis. In particular, sustainable big data analysis requires a cohort of well-supported researchers at senior and junior levels, as well as a pipeline of graduate students and post-doctoral fellows. A crucial consideration is that not only are African science, health, and education budgets relatively small in real terms, but also that those who have posts in

these sectors have other duties (including teaching and clinical duties) that leave little protected time for research (Bezuidenhout et al. 2017). More senior researchers are likely to have significant administrative obligations, while more junior researchers are likely to have significant teaching obligations. Many African countries and institutions have not developed the concept of protected time for specific activities, or alignment of components of salaries with funding sources and proportion of efforts. Additionally, many funding agencies fund only research consumables and very few provide salary support or funds for development of general research infrastructure.

While there is a growing pipeline of postgraduate students in Africa, the flow is not uniform across countries, and it is relatively slow in some regions. Most postgraduate programs do not have postdoctoral components which truncates the training, mentoring and development of African scientists. In fact, there are very few programs with well-developed post-doctoral programs on the continent. Additionally, training of staff and students does not necessarily incorporate training in data management, curation or access. Even if researchers do have easy access to data, in poorly resourced institutions, many lack the capability to make effective use of the data. Therefore it is not only training in access that is important, but also in how to use what can be accessed (Bezuidenhout et al. 2017). H3Africa has provided key resources for data collection and analysis in the short-term, but the longer-term view, in which sustained big data analysis takes place, may well require a different set of resources. An open ecosystem of genomic data requires bioinformatics skills and data curators, positions which may receive low priority for funding in academic institutions. In the absence of sustainable resources for big data analysis in Africa, there is every likelihood that publically shared data will be analyzed by researchers in high-income countries. This likelihood will perpetuate the “research gap”, whereby 90% of the world’s research is done in geographic regions where only 10% of the world’s population lives. This again raises the ethical issues of fairness and equity in sharing, discussed above, but, more importantly, the need for development of genomics skills through training.

Training in genomics data analysis

DNA sequencing is rapidly becoming a core part of medicine for millions of patients due to the decrease in sequencing costs. The data sources being generated by researchers, hospitals, and mobile devices around the world are diverse, complex, disorganized, massive, and of multimodal nature (BD2K, 2016). It is predicted that by 2025, genomics will produce between two and 40 Exabytes of data annually (Stephens et al. 2015). This deluge of data presents new opportunities, as well as new challenges. If researchers can make sense of the wealth of information, there can be an immense advancement in our understanding of human health and diseases. However, apart from the lack of appropriate tools and poor data accessibility, a major barrier to this rapid translational impact is insufficient training in the area of Big Data Analytics in Africa. New bioinformatics curricula can prepare students to address challenges raised by big data in the area of data unification, computational and storage limitations and multiple hypothesis testing (Greene et al. 2016). Data unification refers to the challenges of addressing the inconsistency in data to obtain the necessary data in the appropriate format, as well as normalizing them to make them comparable across sources. This is essential for effective data sharing. H3ABioNet has run short courses on data management and analysis and has developed a recommended curriculum and guidelines for developing new degree programs in bioinformatics (https://training.h3abionet.org/curriculum_development_wg/). These were used for establishing a new Masters degree at the University of Bamako, Mali, which has completed its first successful program this year (2017).

The field is moving rapidly, and the challenges and thus the solutions will keep changing. Scientists with skills in big data will need to be able to understand the current computing environment (processor, storage, memory and network costs) and how to, within that environment, most effectively mine the large-scale data to derive interesting insights. Significant resources are being allocated for training scientists in the analysis of large-scale data in the US (United States) and worldwide. In Africa, various training courses on big data have been provided to scientists and students under both the Square Kilometer Array Africa (www.ska.ac.za/) and IBM MEA (Middle East and Africa) initiatives (www.ibm.com/services/weblectures/meapwww.ibm.com/services/weblectures/meap). IBM has partnered with academia in Africa to promote big data and analytics training programs through its MEA program. In another example, the United Genomes Project (<http://www.unitedgenomes.org/>) aims at helping researchers to enhance genomic medicine by (1) developing methods to assemble genomic data across multiple African ethnicities, (2) building capacity across the continent and (3) facilitating scientific discovery through crowdsourcing and open innovation (Siwo et al. 2015). The United Genomes Project collaborates with existing educational programs at universities or with programs offered by projects such as H3Africa.

A few organizations in Africa, namely H3ABioNet nodes, the African Society for Bioinformatics and Computational Biology, the African Society for Human Genetics, and Teaching and Research in Natural Sciences for Development in Africa (TReND in Africa; <http://trendinafrica.org>) are actively involved in offering training programs in genomics and big data for African scientists (Karikari & Aleksic 2015). The Wellcome Trust also regularly runs genomics-related courses in Africa in their extramural program. H3ABioNet has run a number of courses across all areas of bioinformatics and systems administration to prepare researchers for the deluge of genomics data. The African Genomic Medicine Training Initiative, a spin-off from H3ABioNet and other initiatives has designed and run Genomic Medicine training for African-based healthcare professionals based on collaboratively developed curricula.

Overcoming obstacles/challenges to facilitate translational genomics in Africa

Research output and authorship can serve as a veritable proxy for assessing research capacity of universities and research institutions. A recent study used evidence from genomics publications across Sub-Saharan Africa (SSA) to assess the genomic epidemiology research capacity of scientists in the region from 2004 to 2013. Significant disparities currently exist among SSA countries in genomics research capacity (Adedokun et al. 2016). South Africa has the highest genomics research output, which is reflected in the investments made in its genomics and biotechnology sector (Hardy et al. 2008; Warnich et al. 2011; Ndimba and Thomas 2008; Motari et al. 2004). The study findings call for African governments to increase their investments in building local capacity, provide a sustainable research environment, and encourage joint genomics research among those affiliated with SSA universities. Genomics has a huge potential to improve diagnosis and treatment of several medical conditions in Africa. Integrated translational research inclusive of basic functional genomics could add value to the genomics data generated. While the focus is currently on improving capacity for storage and transfer of genomics data, its interpretation is a critical step towards research and development of genomic products and personalized medicine and to understand population diversity in health and disease. However, infrastructure and scientists to conduct genomics research in the region are still suboptimal. Although the H3Africa efforts hold great promise for the transformation of genomics research in Africa through capacity building and better research facilities, there is a need to document the state of local or regional genomics research productivity in order to guide the equitable distribution of resources (Adedokun et al. 2016). Foreign research investments such as those made by the NIH and Wellcome Trust through H3Africa have given genomics research in Africa a major boost, but the funding is not infinite and research groups in African countries need to work towards long term research sustainability.

As outlined above, one key challenge in data sharing relates to promoting fairness and equity. A crucially important aspect of ensuring fairness lies in empowering African researchers to take intellectual leadership roles in genomic research, which includes leading data analysis. The governance framework plays a critical role in ensuring conditions that allow African researchers a fair chance to analyze their data, first. The H3Africa Data Release Policy incorporates several elements to promote fairness, including for instance a lengthy period during which data will not be released (9 months before submission to EGA is required) and a further twelve month publication embargo for work on certain identified topics after data is released. Whilst the efficacy of ensuring that such conditions promote equitable sharing remains to be proven (Bull 2016), the fact that they were developed with endorsement by both the NIH and the Wellcome Trust is a significant achievement for H3Africa researchers, and sets an important target for other, future genomic research initiatives that seek to use African samples. Since the first dataset from H3Africa has only recently been submitted to the EGA and other projects are just starting to analyze their own data, it is too early to tell what the impact of the policy will be, but the H3Africa consortium has already collectively published more than 70 papers since 2014.

Although some of the other obstacles discussed in this paper cannot be overcome easily, and require improvements in basic infrastructure and service provision by local governments, H3Africa has had a large impact on the development of capacity for genomics research and data generation. H3ABioNet has provided support for overcoming challenges in data transfer, storage and processing, and together with the larger H3Africa consortium, has played a major role in the development of necessary skills. Thus, the consortium is working toward a scenario where genomics data can be generated, stored and analyzed in Africa for the benefit of African scientists and ultimate translation to improve the health of Africans.

Acknowledgements

H3ABioNet is supported by the National Institutes of Health Common Fund (grant number U41HG006941). SNA and CAA are funded by H3Africa's African Collaborative Center for Microbiome and Genomics Research (ACCME) grant supported by the National Institutes of Health (1U54HG006947). OA, OA and MOO are funded by H3Africa's Stroke Investigative Research and Educational Network (U54HG007479). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing Interests

The authors have no competing interests to declare.

References

- Adedokun, B O, Olopade, C O and Olopade, O I** 2016 Building local capacity for genomics research in Africa: recommendations from analysis of publications in Sub-Saharan Africa from 2004 to 2013. *Global Health Action*, 9: 31026. DOI: <https://doi.org/10.3402/gha.v9.31026>
- BD2K** <https://datascience.nih.gov> [Accessed, December, 2016].
- Bezuidenhout, L M, Leonelli, S, Kelly, A H and Rappert, B** 2017 Beyond the digital divide: Towards a situated approach to open data. *Science and Public Policy*, 44(4): 464–475. DOI: <https://doi.org/10.1093/scipol/scw036>
- Bull, S** 2016 Ensuring global equity in open research. London: Wellcome Trust. DOI: <https://doi.org/10.6084/m9.figshare.4055181.v1>
- Bull, S, Cheah, P Y, Denny, S, Jao, I, Marsh, V, Merson, L, Shah More, N, Nhan, L N T, Osrin, D, Tangseefa, D, Wassenaar, D and Parker, M** 2015a Best Practices for Ethical Sharing of Individual-Level Health Research Data From Low- and Middle-Income Settings. *Journal of Empirical Research on Human Research Ethics*, 10: 302–313. DOI: <https://doi.org/10.1177/1556264615594606>
- Bull, S, Roberts, N and Parker, M** 2015b Views of Ethical Best Practices in Sharing Individual-Level Data From Medical and Public Health Research: A Systematic Scoping Review. *Journal of Empirical Research on Human Research Ethics*, 10: 225–38. DOI: <https://doi.org/10.1177/1556264615594767>
- Chokshi, D A, Parker, M and Kwiatkowski, D P** 2006 Data sharing and intellectual property in a genomic epidemiology network: policies for large-scale research collaboration. *Bulletin of the World Health Organisation*, 84(5): 382–7. DOI: <https://doi.org/10.2471/BLT.06.029843>
- Chokshi, D A, Thera, M A, Parker, M, Diakite, M, Makani, J, Kwiatkowski, D P and Doumbo, O K** 2007 Valid consent for genetic epidemiology in developing countries. *PLOS Medicine*, 4(4): e95. 636–41.
- Coleman, C H and Bouësseau, M-C** 2008 How do we know that research ethics committees are really working? The neglected role of outcomes assessment in research ethics review. *BMC medical ethics*, 9: 6. DOI: <https://doi.org/10.1186/1472-6939-9-6>
- De Vries, J, Abayomi, A, Littler, K, Madden, E, McCurdy, S, Ouwe Missi Oukem-Boyer, O, Seeley, J, Staunton, C, Tangwa, G, Tindana, P, Troyer, J and The H3Africa Working Group on Ethics** 2015a Addressing ethical issues in H3Africa research – the views of research ethics committee members. *The HUGO Journal*, 9: 1. DOI: <https://doi.org/10.1186/s11568-015-0006-6>
- De Vries, J, Abayomi, A, Littler, K, Madden, E, McCurdy, S, Ouwe Missi Oukem-Boyer, O, Seeley, J, Staunton, C, Tangwa, G, Tindana, P, Troyer, J and The H3Africa Working Group on Ethics** 2015c Addressing ethical issues in H3Africa research – the views of research ethics committee members. *The HUGO Journal*, 9(1): 1. DOI: <https://doi.org/10.1186/s11568-015-0006-6>
- De Vries, J, Bull, S J, Doumbo, O, Ibrahim, M, Mercereau-Puijalon, O, Kwiatkowski, D and Parker, M** 2011 Ethical issues in human genomics research in developing countries. *BMC Medical Ethics*, 12: 5. DOI: <https://doi.org/10.1186/1472-6939-12-5>
- De Vries, J, Littler, K, Matimba, A, McCurdy, S, Ouwe Missi Oukem-Boyer, O, Seeley, J and Tindana, P** 2016 Evolving perspectives on broad consent for genomics research and biobanking in Africa. Report of the Second H3Africa Ethics Consultation Meeting, 11th May 2015. *Global Health, Epidemiology and Genomics*, 1: 1–3. DOI: <https://doi.org/10.1017/ghg.2016.5>
- De Vries, J, Munung, S N, Matimba, A, McCurdy, S, Ouwe Missi Oukem-Boyer, O, Staunton, C, Yakubu, A, Tindana, P and H3Africa Consortium** 2017 Regulation of genomic and biobanking research in Africa: a content analysis of ethics guidelines, policies and procedures from 22 African countries. *BMC Med Ethics*, 18(1): 8. DOI: <https://doi.org/10.1186/s12910-016-0165-6>

- De Vries, J, Tindana, P, Littler, K, Ramsay, M, Rotimi, C, Abayomi, A, Mulder, N and Mayosi, B M** 2015b The H3Africa policy framework: negotiating fairness in genomics. *Trends in Genetics*, 31: 117–9. DOI: <https://doi.org/10.1016/j.tig.2014.11.004>
- De Vries, J, Williams, T, Bojang, K, Kwiatkowski, D, Fitzpatrick, R and Parker, M** 2014 Knowing who to trust: exploring the role of 'ethical metadata' in mediating risk of harm in collaborative genomics research in Africa. *BMC Medical Ethics*, 15: 62. DOI: <https://doi.org/10.1186/1472-6939-15-62>
- Federal Ministry of Health** 2007 National Code of Health Research Ethics.
- Ferguson, A R, Nielson, J L, Cragin, M H, Bandrowski, A E and Martone, M E** 2014 Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nature Neuroscience*, 17(11): 1442–7. DOI: <https://doi.org/10.1038/nn.3838>
- Foster, I** 2011 Globus Online: Accelerating and Democratizing Science through Cloud-Based Services. *Internet Computing, IEEE*, 15(3): 70–3. <http://doi.ieeecomputersociety.org/10.1109/MIC.2011.64>.
- Gaheen, S, Hinkal, G W, Morris, S A, Lijowski, M, Heiskanen, M and Klemm, J D** 2013 caNanoLab: data sharing to expedite the use of nanotechnology in biomedicine. *Computational Science and Discovery*, 6(1): 014010. DOI: <https://doi.org/10.1088/1749-4699/6/1/014010>
- Goodman, S N** 2015 Clinical trial data sharing: what do we do now? *Annals of Internal Medicine*, 162(4): 308–9. DOI: <https://doi.org/10.7326/M15-0021>
- Goodwin, S, McPherson, J D and McCombie, W R** 2016 Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6): 333–51. DOI: <https://doi.org/10.1038/nrg.2016.49>
- Greene, A C, Giffin, K A, Greene, C S and Moore, J H** 2016 Adapting bioinformatics curricula for big data. *Briefings in Bioinformatics*, 17(1): 43–50. DOI: <https://doi.org/10.1093/bib/bbv018>
- Gymrek, M, McGuire, A L, Golan, D, Halperin, E and Erlich, Y** 2013 Identifying personal genomes by surname inference. *Science*, 339: 321–4. DOI: <https://doi.org/10.1126/science.1229566>
- Hardy, B J, Seguin, B, Ramesar, R, Singer, P A and Daar, A S** 2008 South Africa: from species cradle to genomic applications. *Nature Reviews Genetics*, 9: S19–23. DOI: <https://doi.org/10.1038/nrg2441>
- Homer, N, Szelinger, S, Redman, M, Duggan, D, Tembe, W, Muehling, J, Pearson, J V, Stephan, D A, Nelson, S F and Craig, D W** 2008 Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4: e1000167. DOI: <https://doi.org/10.1371/journal.pgen.1000167>
- Ikingura, J K B, Kruger, M and Zeleke, W** 2007 Health research ethics review and needs of institutional ethics committee in Tanzania. *Tanzania Health Research Bulletin*, 9(3): 154–8.
- Jenkins, C, Arulogun, O S, Singh, A, Mande, A T, Ajayi, E, Benedict, C T, Ovbiagele, B, Lackland, D T, Sarfo, F S, Akinyemi, R, Akpalu, A, Obiako, R, Melikam, E S, Laryea, R, Shidali, V, Sagoe, K, Ibinaiye, P, Fakunle, A G, Owolabi, L F, Owolabi, M O and The SIREN team** 2016 Stroke Investigative Research and Education Network: Community Engagement and outreach within Phenomic core. *Health Education & Behavior*, 43(1S): 82S–92S. <http://journals.sagepub.com/home/heb>. DOI: <https://doi.org/10.1177/1090198116634082>
- Karikari, T K and Aleksic, J** 2015 Neurogenomics: An opportunity to integrate neuroscience, genomics and bioinformatics research in Africa. *Applied and Translational Genomics*, 5: 3–10. DOI: <https://doi.org/10.1016/j.atg.2015.06.004>.
- Kass, N E, Hyder, A A, Ajuwon, A, Appiah-Poku, J, Barsdorf, N, Elsayed, D E, Mokhachane, M, Mupenda, B, Ndebele, P, Ndossi, G, Sikateyo, B, Tangwa, G and Tindana, P** 2007a The structure and function of research ethics committees in Africa: A case study. *PLoS Medicine*, 4(1): e3. DOI: <https://doi.org/10.1371/journal.pmed.0040003>
- Kirigia, J M, Wambebe, C and Baba-Moussa, A** 2005 Status of national research bioethics committees in the WHO African region. *BMC Medical Ethics*, 6: E10. DOI: <https://doi.org/10.1186/1472-6939-6-10>
- Lakes, K D, Vaughan, E, Pham, J, Tran, T, Jones, M, Baker, D, Swanson, J M and Olshansky, E** 2014 Community member and faith leader perspectives on the process of building trusting relationships between communities and researchers. *Clinical and Translational Science*, 7(1): 20–8. DOI: <https://doi.org/10.1111/cts.12136>
- Lambert, C A and Tishkoff, S A** 2009 Genetic structure in African populations: Implications for human demographic history. *Cold Spring Harb Symp Quant Biol*, 74: 395–402. DOI: <https://doi.org/10.1101/sqb.2009.74.053>
- Lemmens, T** 2015 Informed consent. In: Joly, Y and Knoppers, B M (eds.), *Routledge handbook of medical law and ethics*, 27–51. Abingdon, UK, Routledge.


- Marshall, P, Adebamowo, C, Adeyemo, A, Ogundiran, T, Strenski, T, Zhou, J and Rotimi, C** 2014 Voluntary participation and comprehension of informed consent in a genetic epidemiological study of breast cancer in Nigeria. *BMC Medical Ethics*, 15: 38. DOI: <https://doi.org/10.1186/1472-6939-15-38>
- Marzouk, D, Abd El Aal, W, Saleh, A, Sleem, H, Khyatti, M, Mazini, L, Hemminki, K and Anwar, W A** 2014 Overview on health research ethics in Egypt and North Africa. *European Journal of Public Health*, 24(Suppl 1): 87–91. DOI: <https://doi.org/10.1093/eurpub/cku110>
- McGuire, A L and Beskow, L M** 2010 Informed consent in genomics and genetic research. *Annual Review of Genomics and Human Genetics*, 11: 361. DOI: <https://doi.org/10.1146/annurev-genom-082509-141711>
- McWilliams, R, Hoover-Fong, J, Hamosh, A, Beck, S, Beaty, T and Cutting, G** 2003 Problematic variation in local institutional review of a multicenter genetic epidemiology study. *Jama*, 290(3): 360–6. DOI: <https://doi.org/10.1001/jama.290.3.360>
- Miller, GW** 2015 Data sharing in toxicology: beyond show and tell. *Toxicological Sciences*, 143(1): 3–5. DOI: <https://doi.org/10.1093/toxsci/kfu237>
- Motari, M, Quach, U, Thorsteinsdottir, H, Martin, D K, Daar, A S and Singer, P A** 2004 South Africa blazing a trail for African biotechnology. *Nature Biotechnology*, 22: DC37–42. DOI: <https://doi.org/10.1038/nbt1204supp-DC37>
- Mulder, N J, Adebisi, E, Alami, R, Benkahla, A, Brandful, J, Doumbia, S, et al.** 2015 H3ABioNet, a Sustainable Pan African Bioinformatics Network for Human Heredity and Health in Africa. *Genome Research*, 26(2): 271–7. DOI: <https://doi.org/10.1101/gr.196295>
- Munung, N S, Marshall, P, Campbell, M, Littler, K, Masiye, F, Ouwe-Missi-Oukem-Boyer, O, Seeley, J, Stein, D J, Tindana, P and de Vries, J** 2016 Obtaining informed consent for genomics research in Africa: analysis of H3Africa consent documents. *Journal of Medical Ethics*, 42(2): 132–7. DOI: <https://doi.org/10.1136/medethics-2015-102796>
- Ndimba, B K and Thomas, L A** 2008 Proteomics in South Africa: Current status, challenges and prospects. *Biotechnology Journal*, 3: 1368–74. DOI: <https://doi.org/10.1002/biot.200800236>
- Nyika, A, Kilama, W, Chilengi, R, Tangwa, G, Tindana, P, Ndebele, P and Ikingura, J** 2009 Composition, training needs and independence of ethics review committees across Africa: are the gate-keepers rising to the emerging challenges? *Journal of Medical Ethics*, 35(3): 189–93. DOI: <https://doi.org/10.1136/jme.2008.025189>
- Open Knowledge Foundation** The Open Definition. <http://opendefinition.org/> [Accessed, December, 2016].
- Pisani, E and AbouZahr, C** 2010 Sharing health data: good intentions are not enough. *Bulletin of the World Health Organisation*, 88(6): 462–6. DOI: <https://doi.org/10.2471/BLT.09.074393>
- Poldrack, R A and Gorgolewski, K J** 2014 Making big data open: data sharing in neuroimaging. *Nature Neuroscience*, 17(11): 1510–7. DOI: <https://doi.org/10.1038/nn.3818>
- Ramsay, M, De Vries, J, Soodyall, H, Norris, S and Sankoh, O** 2014 Ethical issues in genomic research on the African continent: experiences and challenges to ethics review committees. *Human Genomics*, 8: 15. DOI: <https://doi.org/10.1186/s40246-014-0015-x>
- Ramsay, M, Tiemessen, C T, Choudhury, A and Soodyall, H** 2011 Africa: the next frontier for human disease gene discovery? *Human Molecular Genetics*, 20: R214–20. DOI: <https://doi.org/10.1093/hmg/ddr401>
- Robinson, P N** 2014 Genomic data sharing for translational research and diagnostics. *Genome Medicine*, 6(9): 78. DOI: <https://doi.org/10.1186/s13073-014-0078-2>
- Rotimi, C, Leppert, M, Matsuda, I, Zeng, C, Zhang, H, Adebamowo, C, Ajayi, I, Aniagwu, T, Dixon, M, Fukushima, Y, Macer, D, Marshall, P, Nkwodimmah, C, Peiffer, A, Royal, C, Suda, E, Zhao, H, Wang, VO, McEwen, J and International HapMap Consortium** 2007 Community engagement and informed consent in the International HapMap project. *Community Genetics*, 10(3): 186–98. DOI: <https://doi.org/10.1159/000101761>
- Rung, J and Brazma, A** 2013 Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, 14(2): 89–99. DOI: <https://doi.org/10.1038/nrg3394>
- Schuster, S C, Miller, W, Ratan, A, Tomsho, L P, Giardine, B, Kasson, L R, et al.** 2010 Complete Khoisan and Bantu genomes from southern Africa. *Nature*, 463(7283): 943–7. DOI: <https://doi.org/10.1038/nature08795>
- Shabani, M, Thorogood, A and Borry, P** 2016 Who should have access to genomic data and how should they be held accountable? Perspectives of Data Access Committee members and experts. *European Journal of Human Genetics*, 24: 1671–5. DOI: <https://doi.org/10.1038/ejhg.2016.111>

- Shabrawy, E, El Hefnawy, T and Reda, H** 2014 Applying Ethical Guidelines in Clinical Researches among Academic Medical Staff: An Experience from South Egypt. *British Journal of Medicine and Medical Research*, 4(10): 2014–24. DOI: <https://doi.org/10.9734/BJMMR/2014/7071>
- Siwo, G H, Williams, S M and Moore, J H** 2015 The future of genomic medicine education in Africa. *Genome Medicine*, 7(1): 47. <http://genomemedicine.com/content/7/1/47>. DOI: <https://doi.org/10.1186/s13073-015-0175-x>
- Staunton, C and Moodley, K** 2013 Challenges in biobank governance in Sub-Saharan Africa. *BMC medical Ethics*, 14(1): 35. DOI: <https://doi.org/10.1186/1472-6939-14-35>
- Stephens, Z D, Lee, S Y, Faghri, F, Campbell, R H, Zhai, C, Efron, M J, et al.** 2015 Big Data: Astronomical or Genomical? *PLoS Biology* 13(7): e1002195. DOI: <https://doi.org/10.1371/journal.pbio.1002195>
- Tenopir, C, Allard, S, Douglass, K, Aydinoglu, A U, Wu, L, Read, E, Manoff, M and Frame, M** 2011 Data Sharing by Scientists: Practices and Perceptions. *PLoS One*, 6(6): e21101. DOI: <https://doi.org/10.1371/journal.pone.0021101>
- The H3Africa Consortium** 2014 Enabling African Scientists to Engage Fully in the Genomic Revolution. *Science*, 344(6190): 1346–8. DOI: <https://doi.org/10.1126/science.1251546>
- Tindana, P, Bull, S, Amenga-Etego, L, de Vries, J, Aborigo, R, Koram, K, Kwiatkowski, D and Parker, M** 2012 Seeking consent to genetic and genomic research in a rural Ghanaian setting: a qualitative study of the MalariaGEN experience. *BMC Medical Ethics*, 13(1): 15. DOI: <https://doi.org/10.1186/1472-6939-13-15>
- Tindana, P and De Vries, J** 2016 Broad Consent for Genomic Research and Biobanking: Perspectives from Low- and Middle-Income Countries. *Annual Review of Genomics and Human Genetics*, 17: 2.1–2.19. DOI: <https://doi.org/10.1146/annurev-genom-083115-022456>
- Tindana, P, De Vries, J, Campbell, M, Littler, K, Seeley, J, Marshall, P, Troyer, J, Ogundipe, M, Alibu, V P, Yakubu, A, Parker, M and as members of the H3A Working Group on Ethics** 2015 Community engagement strategies for genomic studies in Africa: a review of the literature. *BMC Medical Ethics*, 16: 24. DOI: <https://doi.org/10.1186/s12910-015-0014-z>
- Tindana, P, Molyneux, C S, Bull, S and Parker, M** 2014 Ethical issues in the export, storage and reuse of human biological samples in biomedical research: perspectives of key stakeholders in Ghana and Kenya. *BMC Medical Ethics*, 15(76): 1–11. DOI: <https://doi.org/10.1186/1472-6939-15-76>
- Tindana, P O, Rozmovits, L, Boulanger, R F, Bandewar, S V, Aborigo, R A, Hodgson, A V, Kolopack, P and Lavery, J V** 2011 Aligning community engagement with traditional authority structures in global health research: A case study from northern Ghana. *American Journal of Public Health*, 101(10): 1857–67. DOI: <https://doi.org/10.2105/AJPH.2011.300203>
- Tishkoff, S A, Reed, F A, Friedlaender, F R, Ehret, C, Ranciaro, A, Froment, A, et al.** 2009 The genetic structure and history of Africans and African Americans. *Science*, 324(5930): 1035–44. DOI: <https://doi.org/10.1126/science.1172257>
- Traore, K, Bull, S, Niare, A, Konate, S, Thera, M A, Kwiatkowski, D, Parker, M and Doumbo, O K** 2015 Understandings of genomic research in developing countries: a qualitative study of the views of MalariaGEN participants in Mali. *BMC Medical Ethics*, 16(42): 42. DOI: <https://doi.org/10.1186/s12910-015-0035-7>
- Van Horn, J D and Ball, C A** 2008 Domain-Specific Data Sharing in Neuroscience: What do we have to learn from each other? *Neuroinformatics*, 6(2): 117–21. DOI: <https://doi.org/10.1007/s12021-008-9019-9>
- van Panhuis, W G, Paul, P, Emerson, C, Grefenstette, J, Wilder, R, Herbst, A J, Heymann, D and Burke, D S** 2014 A systematic review of barriers to data sharing in public health. *BMC Public Health*, 14: 1144. DOI: <https://doi.org/10.1186/1471-2458-14-1144>
- Warnich, L, Drogemoller, B I, Pepper, M S, Dandara, C and Wright, E B** 2011 Pharmacogenomic research in South Africa: Lessons learned and future opportunities in the Rainbow Nation. *Current Pharmacogenomics Personalised Medicine*, 9: 191–207. DOI: <https://doi.org/10.2174/187569211796957575>
- Wright, G E B, Adeyemo, A A and Tiffin, N** 2014 Informed consent and ethical re-use of African genomic data. *Human Genomics*, 8(18): 1–3. DOI: <https://doi.org/10.1186/s40246-014-0018-7>
- Wright, G E B, Koornhof, P G, Adeyemo, A A and Tiffin, N** 2013 Ethical and legal implications of whole genome and whole exome sequencing in African populations. *BMC Medical Ethics*, 14(1): 21. DOI: <https://doi.org/10.1186/1472-6939-14-21>

How to cite this article: Mulder, N, Adebamowo, C A, Adebamowo, S N, Adebayo, O, Adeleye, O, Alibi, M, Baichoo, S, Benkahla, A, Fadlemola, F M, Ghazal, H, Ghedira, K, Matimba, A, Moussa, A, Mungloo-Dilmohamud, Z, Owolabi, M O, Radouani, F, Rotimi, C N, Stein, D J and Souiai, O 2017 Genomic Research Data Generation, Analysis and Sharing – Challenges in the African Setting. *Data Science Journal*, 16: 49, pp.1–15, DOI: <https://doi.org/10.5334/dsj-2017-049>

Submitted: 11 January 2017 **Accepted:** 16 October 2017 **Published:** 03 November 2017

Copyright: © 2017 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 Ubiquity Press.

OPEN ACCESS 