Development of Bioinformatics Infrastructure for Genomics Research

Nicola J. Mulder*, Ezekiel Adebiyi[†,‡], Marion Adebiyi[†,‡], Seun Adeyemi[‡,§], Azza Ahmed[‖], Rehab Ahmed[‖], Bola Akanle[‡,§], Mohamed Alibi[¶], Don L. Armstrong[#], Shaun Aron**, Efejiro Ashano[‡,††], Shakuntala Baichoo[‡‡], Alia Benkahla[¶], David K. Brown[§§], Emile R. Chimusa*[,‖‖], Faisal M. Fadlelmola[‖,¶¶], Dare Falola[‡], Segun Fatumo[††], Kais Ghedira[¶], Amel Ghouila[##], Scott Hazelhurst**, Itunuoluwa Isewon[†,‡], Segun Jung***, Samar Kamal Kassim[†††], Jonathan K. Kayondo[‡‡‡], Mamana Mbiyavanga*, Ayton Meintjes*, Somia Mohammed[‖], Abayomi Mosaku[‡], Ahmed Moussa[§§§], Mustafa Muhammd[‖], Zahra Mungloo-Dilmohamud[‡‡], Oyekanmi Nashiru[††], Trust Odia[‡], Adaobi Okafor[‡], Olaleye Oladipo[‡,‖‖‖], Victor Osamor[†,‡], Jellili Oyelade[†,‡], Khalid Sadki[¶¶¶], Samson Pandam Salifu[###,****], Jumoke Soyemi[††††], Sumir Panji*, Fouzia Radouani[‡‡‡‡], Oussama Souiai[¶], Özlem Tastan Bishop[§§] : and The H3ABioNet Consortium, as members of the H3Africa Consortium

Cape Town, South Africa; Ota, Nigeria; Khartoum, Sudan; Tunis, Tunisia; Champaign, IL, USA; Johannesburg, South Africa; Abuja, Nigeria; Grahamstown, South Africa; Tunis-Belvédère, Tunisia; Chicago, IL, USA; Cairo, Egypt; Entebbe, Uganda; Tangier, Morocco; Omu-Aran, Nigeria; Rabat, Morocco; Kumasi, Ghana; Ilaro, Nigeria; and Casablanca, Morocco

ABSTRACT

Background: Although pockets of bioinformatics excellence have developed in Africa, generally, large-scale genomic data analysis has been limited by the availability of expertise and infrastructure. H3ABioNet, a pan-African bioinformatics network, was established to build capacity specifically to enable H3Africa (Human Heredity and Health in Africa) researchers to analyze their data in Africa. Since the inception of the H3Africa initiative, H3ABioNet's role has evolved in response to changing needs from the consortium and the African bioinformatics community.

Objectives: H3ABioNet set out to develop core bioinformatics infrastructure and capacity for genomics research in various aspects of data collection, transfer, storage, and analysis.

Methods and Results: Various resources have been developed to address genomic data management and analysis needs of H3Africa researchers and other scientific communities on the continent. NetMap was developed and used to build an accurate picture of network performance within Africa and between Africa and the rest of the world, and Globus Online has been rolled out to facilitate data transfer. A participant recruitment database was developed to monitor participant enrollment, and data is being harmonized through the use of ontologies and controlled vocabularies. The standardized metadata will be integrated to provide a search facility for H3Africa data and biospecimens. Because H3Africa projects are generating large-scale genomic data, facilities for analysis and interpretation are critical. H3ABioNet is implementing several data analysis platforms that provide a large range of bioinformatics tools or workflows, such as Galaxy, the Job Management System, and eBiokits. A set of reproducible, portable, and cloud-scalable pipelines to support the multiple H3Africa data types are also being developed and dockerized to enable execution on multiple computing infrastructures. In addition, new tools have been developed for analysis of the uniquely divergent African data and for downstream interpretation of prioritized variants. To provide support for these and other bioinformatics queries, an online bioinformatics helpdesk backed by broad consortium expertise has been established. Further support is provided by means of various modes of bioinformatics training.

Conclusions: For the past 4 years, the development of infrastructure support and human capacity through H3ABioNet, have significantly contributed to the establishment of African scientific networks, data analysis facilities, and training programs. Here, we describe the infrastructure and how it has affected genomics and bioinformatics research in Africa.

The authors report no relationships that could be construed as a conflict of interest.
H3ABioNet is supported by the National Institutes of Health Common Fund (grant number U41HG006941).
The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
From the *Computational Biology Division, Department of Integrative Biomedical Sciences, Institute for Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa; †Department of Computer and Information Sciences, Covenant University, Ota, Nigeria; ‡Covenant University Bioinformatics Research (CUBRe), Covenant University, Ota, Nigeria; §Center for System and Information Service, Covenant University, Ota, Nigeria; ‖Centre for Bioinformatics and Systems Biology, Faculty of Science, University of Khartoum, Khartoum, Sudan; ¶Laboratory of Bioinformatics, Biomathematics and Biostatistics (BIMS), Institut Pasteur de Tunis, Tunis, Tunisia; #Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Champaign, IL, USA; **Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa; ††H3Africa Bioinformatics Network (H3ABioNet) Node, National Biotechnology Development Agency (NABDA), Federal Ministry of Science and Technology (FMST), Abuja, Nigeria; ‡‡University of Mauritius, Moka, Mauritius; §§Research Unit in Bioinformatics (RUBi), Department of Biochemistry and Microbiology, Rhodes University,

Grahamstown, South Africa; |||Division of Human Genetics, Department of Pathology, Faculty of Health Sciences, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa; ¶¶Future University of Sudan, Khartoum, Sudan; ##Institut Pasteur de Tunis, LR11IPT02, Laboratory of Transmission, Control and Immunobiology of Infections (LTCII), Tunis-Belvédère, Tunisia; ***Computation Institute, University of Chicago and Argonne National Laboratory, Chicago, IL, USA; †††Medical Biochemistry and Molecular Biology Department, Faculty of Medicine, Ain Shams University, Abbaseya, Cairo, Egypt; ‡‡‡Uganda Virus Research Institute (UVRI), Entebbe, Uganda; §§§LAbTIC Laboratory, ENSA, Abdelmalek Essaadi University, Tangier, Morocco; ||||Center for System and Information Service, Landmark University, Omu-Aran, Nigeria; ¶¶¶School of Sciences, Mohammed V University of Rabat, Rabat, Morocco; ###Department of Biochemistry and Biotechnology, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana; ****Kumasi Centre for Collaborative Research, South End Asougya Road, KNUST Campus, Kumasi, Ghana; ††††Department of Computer Science, Ilaro Polytechnic, Ilaro, Nigeria; and the ‡‡‡‡Chlamydiae and Mycoplasma Laboratory, Institut Pasteur du Maroc, Casablanca, Morocco. Correspondence: N.J. Mulder (Nicola.mulder@uct.ac.za).

GLOBAL HEART

VOL. 12, NO. 2, 2017
ISSN 2211-8160/$36.00.
http://dx.doi.org/10.1016/j.gheart.2017.01.005

Africa is currently undergoing an epidemiological transition, with endemic infection and a rapidly growing burden of cardiometabolic and other noncommunicable diseases. Understanding the genetic determinants of diseases can lead to novel insights into disease etiology, which may identify novel therapeutic targets, and the potential for better disease prognosis and management. Genomics research holds great promise for medical and health care research and is gaining global momentum as we transition to an era of precision medicine, whereby treatment of individual patients is driven by a greater understanding of the clinical diagnosis through interpretation of underlying genomic variation. With sequencing costs dropping, whole genome sequencing as an aid to diagnosis is becoming more affordable. However, this does not consider the hidden costs required for analysis and interpretation of the data, which is substantial [1]. The decreasing costs associated with next-generation sequencing (NGS) technologies have been accompanied by increasing size and complexity of the sequence data. To deal with such complex and voluminous data, existing data storage and transfer mechanisms as well as public repositories and data processing technologies have had to adapt, and skills in data science have had to be developed [1]. As a consequence, bioinformaticians have become essential to biomedical research projects. This is also true of other large-scale technologies, such as genotyping by arrays, which may not generate quite the same data sizes as NGS, but with data for millions of single nucleotide polymorphisms (SNPs) being generated and analyzed, the processing and downstream analysis require substantial computing resources and associated skills. Muir et al. [1] describe 4 key adaptations that have been required for embracing the genomics era, particularly in relation to NGS data: development of algorithms to handle short reads and long reference genomes, new compression formats for facilitating efficient data storage, adoption of distributed and parallel computing, and increased data security protocols. Bioinformaticians with data science skills are essential for processing, analyzing, and integrating genomics data, but there is also a need to train geneticists and clinicians to interpret and translate the results.

Although large-scale genomics projects have been undertaken for several years internationally, Africa has lagged behind due to limited infrastructure for implementing such large projects. Initiatives such as the H3Africa (Human Heredity and Health in Africa) (http://h3africa.org/) [2] are accelerating genomics research on the continent by funding research as well as building capacity and infrastructure. One component of the H3Africa initiative is H3ABioNet [3], a pan-African bioinformatics network for H3Africa (http://www.h3abionet.org), which is building capacity on the bioinformatics front to enable genomics research on the continent. Here, we describe how this bioinformatics capacity development has been undertaken, demonstrated with examples and the potential effect on genomics research.

APPROACH AND IMPLEMENTATION

H3ABioNet is an extensive network covering over 30 institutions in 15 African countries and 2 partners outside of Africa, and it includes a large diversity of skills. However, the task of bioinformatics capacity development in most African countries is large, and due to limited resources, infrastructure, and expertise, requires careful coordination and pooling of efforts. With limited infrastructure in place prior to the H3Africa initiative, the network had to work on all fronts from genomics data management to building capacity for local analysis of the data.

Tools that are being developed and implemented are addressing the diverse bioinformatic needs of multisite clinical research projects. This includes tools for monitoring recruitment of participants, optimization of data transfer between project sites and to public repositories, harmonization of data, and establishment of standard and custom workflows for data analysis. H3ABioNet has made these available via several web-based, easy to use platforms such as REDCap, Galaxy, or WebProtégé. The network also provides access to experts from a variety of domains to address questions about the established infrastructure and to provide support to H3Africa and other genomics projects. The development of bioinformatics capacity and user support undertaken by H3ABioNet can be divided into several categories related to data transfer and storage; data collection, management, and integration; data analysis and development of associated tools; and training on all of these aspects. Further details about these developments are described in the following text.

GENOMICS SKILLS TRAINING

Genomics, being a multidisciplinary field, requires cross-disciplinary skills involving a combination of knowledge and proficiency from the fields of biology, computer science, mathematics, and statistics. There are numerous challenges in the field of genomics, including the major challenge of processing the massive amount of data and extracting biological meaning from them [4]. In view of these challenges, there is an increased demand for highly trained and experienced bioinformatics experts who can handle the data inundation as well as interface with biologists. Thus, at the core of enabling genomic research is the development of human capacity and the promotion of interdisciplinary training. H3ABioNet has undertaken a multifaceted approach to training, including integrating training with other activities, such as webinars, data analysis, or development hackathons, and the inclusion of shadow teams in projects. H3ABioNet delivers formal training through internships (enabling 1-on-1 skills transfer), short specialized courses, hackathons, and online distributed courses. The training has covered various aspects of bioinformatics from general introductory topics to specialized subjects such as NGS and GWAS analyses. Since 2013, H3ABioNet has conducted over 25 bioinformatics courses across Africa (http://h3abionet.org/training-and-education/h3abionet-courses). Follow-up surveys have

been developed and are sent out on a regular basis to all participants who have attended any H3ABioNet training to assess the long-term effect of training and track the career development of young researchers. Individual nodes also conduct their own training programs within their institutions or across their local regions, and some have started new bioinformatics degree programs since the start of the project.

H3ABioNet webinars form part of the regular H3ABioNet activities and help strengthen research activities and foster collaboration amongst the nodes. The inaugural session for the webinar series was launched in May 2015 and has covered a broad range of relevant bioinformatics topics, including GWAS and population genetics, metagenomics, big data, NGS, cloud computing, and reproducible science. The webinars take place monthly with 2 speakers per session and are educational for all participants. Early in the project, H3ABioNet also established a Node Accreditation Exercise with the aim to give nodes the opportunity to demonstrate that they have a reasonable level of technical competence on essential data analysis workflows relevant to H3Africa. The node accreditation exercise helps African scientists to develop technical skills and build infrastructure for a particular data analysis workflow.

DATA TRANSFER AND SECURE STORAGE

Several challenges face big biological data transfer to, within, and out of Africa. These include slow and unstable Internet connectivity; unreliable power supply; continent-wide obsolete computer infrastructure that varies between medium-scale server infrastructure to a small number of workstations, with multiple operating systems (Windows, Linux) for different purposes; and a lack of centralized and secure data storage. Due to the cost and sensitive nature of human genetics data, it was important to establish a system for reliable and secure data transfer in Africa. As a first step, we evaluated the connectivity between different collaborating endpoints, followed by the identification of a high-performance transfer service that allows stable, secure, and synchronized file transfer. The NetMap project was set up and used to build an accurate picture of network performance within Africa and between Africa and the rest of the world. Tools, based on the *iPerf* toolbox (http://iperf.sourceforge.net/), were developed to monitor the performance of network connections. These have been installed at the H3ABioNet nodes, and a visualization dashboard was developed to report Internet speeds between sites daily. Through the NetMap project, we have gained the capacity to gather accurate data about the effective bandwidth of links between the nodes. The results have demonstrated that actual speeds vary immensely from bit/s to Gbits/s and seldom reach the speeds the institutions claim to offer. This information has been used to justify upgrades and fixes to some of the nodes' infrastructure where bandwidth may be limiting.

A technical solution that enables fast, secure, and reliable transfer of large datasets, independently of the bandwidth and quality of Internet connections, was needed for collaborators in Africa. Globus Online (https://www.globus.org/) [5] is a grid-scale file transfer service that offers the right technologies (GridFTP, Third party file transfer control, and so on) to implement a solution to make data transfer in Africa more reliable. Through Globus scalability and some optimizations made especially to the H3ABioNet nodes, it has addressed most of the roadblocks encountered while using traditional file transfer protocols such as FTP or HTTP. The H3ABioNet infrastructure working group (ISWG) developed standard operating procedures (SOPs) for the configuration of hosts running Globus Online, the implementation of the service, and the best practices for using it, either for individual usage or between teams at an institutional level. Following the SOP, many H3ABioNet nodes have implemented Globus Online in their infrastructure, and others have already used its services for internode collaborative work and transparent sharing of research data. As examples, Globus was used to transfer 668,622 individual files from Rhodes University to the Centre for High Performance Computing in Cape Town, and for transferring 140 TB of NGS data from the United States to South Africa. The H3ABioNet ISWG also developed an SOP for the usage of the H3Africa archive architecture (see the following text) as the centralized storage area for H3Africa data. This is linked to an endpoint that enables project data to be transferred using Globus Online services. Addressing the challenges led the ISWG to create a number of publicly available documents with detailed information on how to resolve technical issues to establish an active Globus node.

One of the funder requirements for H3Africa data is the submission to public data repositories in a timely manner. To facilitate this, an H3Africa archive was developed to host and prepare the data for submission to public repositories, and timing of submissions could be monitored. The archive architecture was designed on the basis of the one used by the European Bioinformatics Institute's European genome-phenome Archive (EGA) [6], to ensure data security. The data has to be submitted to the EGA within 9 months of reaching the archive, during which time it is prepared for submission by ensuring that the format and content are EGA compliant. H3ABioNet acts as the data coordinating center that liaises with the EGA and provides support for preparing data for submission.

DATA COLLECTION, MANAGEMENT, AND INTEGRATION

Central to the H3Africa projects is the collection and management of data, which can be of many different types, including clinical, demographic, genomic, and so on. Clinical and specifically phenotype data are collected during participant recruitment and consultations, whereas other data may be collected or generated at multiple sites,

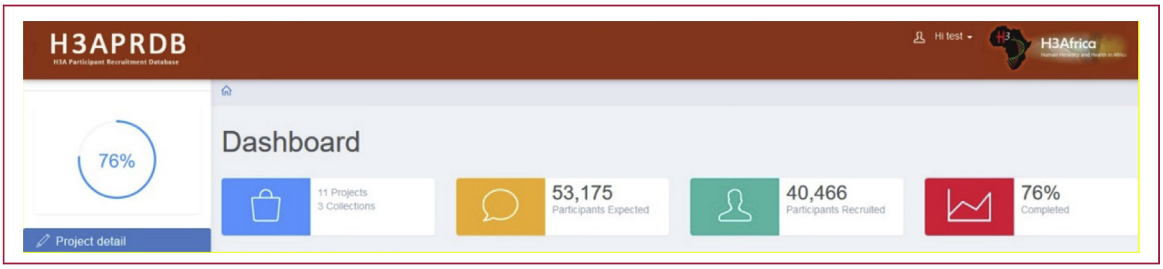

FIGURE 1. H3Africa participant recruitment database quick report section.

including laboratories, sequencing centers, or other core facilities. Whatever the nature of the data, it is essential that it is accurate, complete, and well curated and managed. Data also needs to be tracked and monitored to ensure provenance.

PARTICIPANT RECRUITMENT

Recruitment of the target number of participants is critical for a clinical study. Failure to achieve the targeted enrollment and retention rates can have scientific consequences, as studies can experience reduced statistical power. At the request of the H3Africa Consortium, H3ABioNet designed an H3Africa Participant Recruitment Database (H3APRDB) (https://redcap.h3abionet.org/h3aprdb/). The H3APRDB, backed by REDCap [7], is a data submission and reporting interface designed for the H3Africa principal investigators and funders to track participant recruitment at the site, project, and consortium levels, with built in graphical outputs. In the design, a few of the questions identified for the H3APRDB were common to all projects, whereas some were project specific. Therefore, the data submission interfaces had to be customized for each project, and these were pre-populated with the project sites, as well as expected recruitment numbers. Users access the H3APRDB via the Internet using their respective login credentials, and then the system filters the users on the basis of their unique login parameters and renders the required reporting information on the basis of the role assigned to the user. Each principal investigator only has access to his/her own project information, whereas the funders can access results from any of the projects they fund and get an overview of all funded projects.

In the H3APRDB, REDCap forms are populated quarterly by the projects using the pre-established project questions. It was designed as a server-based application, which makes it accessible remotely via an Internet browser. The user interface was also programmed with mobile responsive features using the Cascading Style Sheet 3 "@media" query, which makes it suitable for any mobile device and tablet. A reporting system was developed allowing for automatic export of collected data from REDCap in JSON format and subsequent processing to generate graphs to demonstrate recruitment progress. The user interface was implemented using Hypertext Mark-up Language and JavaScript, using Cascading Style Sheet for styling, and the reporting system connects to the REDCap application programming interface via server-side components implemented in PHP. The code is available in the GIT hub at: https://github.com/h3abionet/h3aprdb. The reporting interface provides an overview of recruitment progress over time, the percentage completeness by project, and the number of recruited participants per site. Bar charts display the recruitment numbers at each data collection point against the total number to be recruited, enabling a quick comparison of the progress of all projects at a glance. The charts have clickable project titles on the vertical axis, which link to the unique data for that particular project. The dashboard also features active legends that ease navigation between reports, and buttons to hide or display selected project data. The dashboard has a quick report section at the top of the page (Fig. 1), which automatically counts and reports specific numerical values. This provides an up to date count of the number of projects, number of collections, number of expected participants, actual number of participants recruited, and percentage completion over all projects.

The dashboard has built-in JavaScript functions for automatic translation of the reported data and charts into different formats for printing, further use in other software, or sharing generated reports. It auto-generates pdf, png, jpg, or SVG formats of the graphed reports. In providing a single common reporting platform for all of the H3Africa projects, recruitment progress can easily be tracked, allowing for better monitoring of project expectations and trends.

DATA CURATION AND STANDARDIZATION

Harmonizing data across multiple projects or even multiple sites within a project is challenging, but is also necessary to enable cross-consortium analyses. The heterogeneity of the terminologies used for the description of H3Africa metadata made it necessary to adopt ontologies and controlled vocabularies. Ontologies enable unambiguous contextual description of data, as terms have descriptions and relationships between them. Consistent description of data is required before submission of data to the EGA and biospecimens to the biorepositories. This requires adequate sample, phenotype, and experimental procedure description and coordination between data and biospecimen

repositories. Data curation and standardization is ongoing alongside the submission of data to the H3Africa archive and biorepositories. We are using a simple metadata schema that includes publicly accessible data types that will be available from the EGA and biorepository web sites after submission. This schema, and the structure of underlying ontologies to which the data will be mapped, will be used in an H3Africa catalogue to enabling search for specific data or biospecimens.

Where unique data generated by H3Africa projects could not be mapped to existing ontologies, we sought to extend these ontologies to make them more applicable to Africa. For example, H3ABioNet members and the sickle cell disease (SCD) community joined forces to develop an ontology covering aspects related to SCD under the classes: phenotype, diagnostics, therapeutics, quality of life, disease modifiers, and disease stage. Experts in SCD from around the world contributed to the development of the SCD ontology at a hands-on workshop, the proceedings of which were published [8]. We settled on Protégé for an initial implementation of ontology structures, and have used its web interface WebProtégé [9] for the follow-up developments by domain experts. The generic workflow used for building ontologies starts by enumerating the important terms; reviewing existing ontologies; defining classes, subclasses, and the hierarchy between classes; defining the properties of the classes; defining the different facets of the classes; and filling in the values. The SCD ontology is continuing to be developed, curated, and integrated into the WebProtégé platform and the web site dedicated to this project. The final ontology will be reviewed by an international advisory group.

DATA INTEGRATION

There will be many different data types generated by H3Africa projects. These include phenotype, microbiome data, exome, whole genome sequence, and GWAS data. These data types need to be integrated within projects to answer the research questions, but also at a higher level to get a broad picture of data from the consortium. Information about these different data types, or metadata, is essential for understanding, sharing, and integration. The metadata, which includes clinical, demographic, and genomic or other experimental data from projects, will be available for searching via the H3Africa catalogue mentioned previously, but ultimately will be integrated into a more comprehensive African variation database.

Relevant public data is also being integrated into the Human Mutation Analysis (HUMA) resource (https://huma.rubi.ru.ac.za/), which is a platform for the analysis of genetic variation in humans, with the focus being on the downstream analysis of variation in protein sequences and structures. It integrates data from various sources in a single connected database, including UniProt [10], dbSNP [11], Ensembl [12], the Protein Data Bank [13], Human Genome Nomenclature Committee [14], ClinVar [15], and OMIM (Online Mendelian Inheritance in Man) [16]. Data collected includes all human genes, protein sequences and structures, exons and coding sequences, genetic variation, and diseases that could be mapped to variations and genes. This data was integrated into a MySQL database and was made publicly accessible via a user-friendly web interface and RESTful web application programming interface. The web server was developed using the Django web framework. Additionally, HUMA has been designed to allow variation data in VCF format to be uploaded in the form of private datasets. This data is stored separately from the public data and is kept private unless the user explicitly shares it. This data is automatically mapped to protein sequences and structures (using chromosome co-ordinates) already in the database, allowing users to visualize the location of the uploaded variation in 3-dimensional space. Currently, HUMA contains over 22,000 genes, 157,000 unique protein sequences, 14,000 diseases, 32,000 protein structures, and approximately 71 million variants mapped to proteins and genes. HUMA serves as a resource for H3Africa projects, providing a space for unique variation to be uploaded, compared with existing mutations from dbSNP and UniProt, and analyzed computationally.

DATA ANALYSIS

Analysis of large-scale genomic data and integration with phenotypes is not trivial. It requires tools, computing infrastructure, and relevant skills. To support data analysis, H3ABioNet has developed several new tools, developed SOPs for pipelines required for H3Africa data analysis, and explored computing facilities and how to access these. Support for all of these is provided through a ticketed helpdesk.

SOPs AND COMPUTING PLATFORMS

Genomic projects currently funded through H3Africa typically involve the generation of genetic data through genotyping arrays, whole genome or exome sequencing, whereas the microbiome projects are mostly using 16S ribosomal ribonucleic acid sequencing. Therefore, we have developed SOPs for analysis pipelines for these data; the steps to follow are divided into pre-processing and quality control, analysis, and interpretation, with software for each step and approximate computing requirements (central processing units and storage). The SOPs are publicly available from the H3ABioNet website (http://www.h3abionet.org/tools-and-resources/sops), with practice datasets for researchers who want to test the pipelines themselves. The SOPs and datasets are tools for helping the H3ABioNet nodes prepare for Node Accreditation exercises and for others who want to learn about the pipelines.

Data analysis is variable, requiring anything from running a few scripts to complex analysis workflows with data visualization tools. Several data analysis platforms that provide a large range of bioinformatics tools exist. Galaxy (http://usegalaxy.org) [17] is an example, which provides a

user friendly graphical interface to a large range of bioinformatics tools for the analysis of various types of data. Galaxy supports reproducible computational research by providing an environment for performing and recording bioinformatics analyses. Several Galaxy instances have been set up by different H3ABioNet nodes (Tunisia: http://tesla.pasteur.tn:8080/; Morocco: www.ensat.ac.ma/mobihic/Galaxy/; and South Africa [currently private]), and well-documented standard workflows that are useful for the analysis of H3Africa data have been implemented. Nodes have also included specific pipelines relevant to their regions; for example, the Moroccan instance implemented a global workflow for microarray and mass spectroscopy data analysis in their Galaxy platform. The eBiokit (http://www.ebiokit.eu/) is a standalone device that can run bioinformatics analyses independent of the Internet. This is particularly useful for training where Internet access is unreliable, and it also hosts a Galaxy instance.

Another resource to store network-built tools and workflows is the Job Management System (JMS) [18], a cluster front-end and workflow management system designed to ease the burden of using HPC resources and facilitate the sharing of tools and workflows. JMS, developed within H3ABioNet, allows users to upload tools and scripts via a user-friendly web interface or tools can be combined via a drag-and-drop interface, to create complex workflows. JMS automatically generates web-based interfaces to tools and workflows on the basis of user-provided information on how each tool can be executed. Tools and workflows can then be shared between JMS users. JMS is being set up at the Centre for High Performance Computing, where it will be freely available to all South African–based academics. For others, JMS can be downloaded and installed on local infrastructure. Tools and workflows collected within H3ABioNet will be shared across all JMS instances, providing each node with the ability to perform multiple types of computational analysis. Those housed in JMS have also been made public via external web interfaces. Future development of JMS will focus on creating a grid computing network incorporating all groups who have set up a JMS instance. This will facilitate the sharing of computational resources between participating nodes. JMS is open-source and can be downloaded from https://github.com/RUBi-ZA/JMS.

More recently, technologies have developed for packaging tools to deploy on various high-performance computing infrastructures, including Cloud platforms. We are developing a set of simple portable pipelines that scale well to support multiple H3Africa projects. These use popular workflow management languages to describe the workflows (either Common Workflow Language or Nextflow), which allow the specification of robust and reproducible workflows that are fault tolerant and provide the end users with instructions for running the workflow. We have also used Docker, a containerization service that abstracts from the underlying system and hardware to promote portability and ease of installation. Four containerized pipelines are currently under construction. The *H3Agwas* pipeline will support the calling of genotypes, a range of quality control steps and association tests, including linear and logistic regression and mixed model approaches, and will also support population structure analysis. The workflow supports native and dockerized execution on single servers or clusters, and an Amazon AMI is available for execution on Amazon. The second pipeline uses imputation to address the issue of missing information in large-scale data. This Nextflow-based (https://www.nextflow.io/) imputation workflow utilizes IMPUTE2 [19] and SHAPEIT [20] to phase typed genomic variants and impute untyped genomic variants using a previously phased reference panel. To speed up the computation, the workflow parallelizes imputation by chunking across chromosome windows. The workflow is designed to be run either within Docker containers in a Docker swarm deployed on cloud infrastructure (using OpenStack or Amazon AWS) or within a local HPC installation which uses slurm or another scheduler which is supported by Nextflow. The 2 additional workflows process NGS data, one specializing in 16S ribosomal ribonucleic acid data for metagenomic analysis, and the other in exome sequence data, with downstream variant calling. Much of the emphasis on the development of these pipelines has been on portability to take into account the heterogeneous computing environments and infrastructure within Africa to allow groups to run these pipelines on different environments (servers, clusters, high performance computing centers and cloud computing platforms) while also ensuring scalability.

ANALYSIS TOOLS

There is an extensive collection of bioinformatics tools available in the public domain for the analysis of many different data types. However, there is seldom a "1 solution fits all" resource, and sometimes new, unique data brings new analysis challenges. Although it is important not to "reinvent the wheel," there is constantly a need for tool adaptation, customization, or redevelopment to fill gaps in existing tools and pipelines. African genomic data is unique in its novelty and diversity, necessitating new or adapted algorithms for data analysis and visualization. Some examples of new or adapted tools developed by H3ABioNet are described in the following text.

Inference of the genetic ancestries in admixed populations is an important area of study, with applications in admixture mapping (method used to localize disease causing genetic variants which differ in frequency across populations). Several methods have been developed for inferring local ancestry, but many still have limitations in working with complex multiway admixed populations. Due to the historical action of natural selection, the modeled ancestral population in some chromosomal regions may be divergent from the true ancestral population. These spurious deviations would be present in both cases

and control subjects, and would lead to spurious case-only admixture associations or admixture mapping. Therefore, we developed 4 tools related to admixture analysis, including a tool for selecting the best proxy ancestral populations for an admixed population, algorithms for dating different admixture events in multiway admixture populations and for inferring local ancestry in admixed populations, and a tool for mapping genes underlying ethnic differences in complex disease risk in multiway admixed populations. The first tool, PROXYANC [21] is an important precursor for the rest of the tools, as identifying the correct ancestral populations is crucial to enable accurate inference of local ancestry, mapping disease, and dating admixture events in mixed populations. It incorporates 2 novel algorithms, including the correlation between observed linkage disequilibrium in an admixed population and population genetic differentiation in ancestral populations, and an optimal quadratic programming on the basis of the linear combination of population genetic distances. For dating distinct ancient admixture events in a multiway admixed population, we are currently implementing DateMix, a new method that is based on the distribution of ancestry segments along the genome of the admixed individuals. We are also designing a model that accounts for historical gene flow and natural selection to achieve superior accuracy in inferring ancestry of origin at every genomic locus in a multiway admixed population. This new approach, called ancENS, makes use of the Approximate Bayesian computation sampling approach to approximate the posterior probability of the modeled ancestral hidden Markov model. Finally, since the locus-specific ancestry along the genome of an admixed population can boost the power to detect signals of disease genes for complex disease that differ in prevalence among different ethnic groups, we are developing new disease scoring statistics for multiway admixed populations that account for admixture association, gene-environment interactions, and family relationships.

GWA studies on complex diseases often provide multiple significant SNPs or no significant associations but a set of potential SNPs that fall just below the significance threshold. For downstream analysis of variants from a genome-wide association study (GWAS) or NGS experiment, we developed ancGWAS [22] and HUMA. ancGWAS is a network-based tool for analysis of GWAS summary statistics that identifies significant subnetworks in the human protein−protein interaction network by mapping SNPs to genes and genes to pathways and networks. ancGWAS enables one to make sense of how multiple SNPs/genes that are found to be significant in a GWAS may be related to each other and which biological pathways are enriched with these genes. When working with admixed populations, the tool is also able to identify genes or subnetworks that differ in ancestry proportions compared with the average proportions across the genome. Another useful pipeline developed within H3ABioNet is a variant prioritization workflow and associated guidelines. The variant annotation uses existing tools such as ANNOVAR, PolyPhen, and so on, and then a variant prioritization workflow determines which SNPs out of a set of SNPs generated by an experiment should be prioritized for further analysis to interpret the results of the experiment. A final example is HUMA, which enables downstream interpretation by mapping variants to protein structures purely on the basis of the chromosome coordinates and visualizing these variants in the structure. Additionally, HUMA includes Vapor, a variant analysis workflow, which combines the predictions of Provean [23], PolyPhen-2 [24], PhD-SNP [25], FATHMM [26], AUTO-MUTE [27], MuPro [28], and I-Mutant 2.0 to predict the effects of mutations on protein function and stability. HUMA also integrates PRIMO (Protein Interactive Modeling, https://primo.rubi.ru.ac.za/), a homology modeling pipeline [29]. The HUMA web server makes use of JMS to run tools and workflows on the underlying cluster.

HELPDESK

With the development of a variety of tools or computing facilities for addressing many aspects of bioinformatics and genomics research, and also taking cognizance of the fact that many wet-lab researchers have little experience in bioinformatics, H3ABioNet set up an online helpdesk to provide access to expertise in the field. The helpdesk offers technical support in various bioinformatics and genomics categories to H3Africa and non-H3Africa projects in need of assistance with study design and analysis of genomics experiments. Queries are handled by a team of experts from H3ABioNet, which collectively have broad experience ranging from whole genome sequencing, genotyping, and systems administration to software engineering. User requests are categorized and assigned to experts, and the expected turnaround time for dealing with a request is monitored. The helpdesk, accessed via a submission interface available at http://www.h3abionet.org/support, is audited every 6 months to monitor progress and deliver strategies to improve interventions. Users can also browse through the helpdesk knowledge base or additional resources, which include specialized discussion fora and other bioinformatics platforms that could potentially answer their queries.

DISCUSSION AND CONCLUSIONS

Considering the fact that the largest human genetic diversity lies within the African continent, African genomic data will likely provide unique clinical and biodiversity knowledge [30,31]. However, genomic data poses theoretical challenges through the massive data volume, dimensionality of the data, and African-specific statistical approaches required for mining and interpreting the data. These challenges range from the physical aspects of dealing with these data through to the biomedical interpretation for the ultimate improvement of health care. The tools that H3ABioNet has developed are

relatively new, but were developed in response to an immediate need. Therefore, most have been implemented in research projects, some of which are published and some still in progress. Although robust infrastructure remains a central issue in most African countries, access to computers, tools, and training for bioinformatics is being addressed through H3ABioNet, which has contributed significantly to infrastructure and human capacity development in local African nodes within the network. This has the potential to considerably improve the quality of genomics research and enable data analysis by African scientists.

In the past 5 years, there has been a considerable increase in the availability of bioinformatics training programs in many countries [32], which have trained a large number of African-based scientists to develop experts in the discipline. Additionally, improving the computing infrastructure and access to tools across the continent has made large-scale genomic data analysis more accessible for these scientists. At the start of the project, few African groups had the capacity or skills to analyze large genomics datasets; however, there are now several teams who have both the skills and infrastructure to do so, as demonstrated through successful completion of the H3ABioNet node accreditation exercises, and the analysis of over 5,000 full genome sequences for the design of a new genotyping array. The investments in infrastructure support are making critical contributions to advancing biomedical research and associated bioinformatics applications. Networking through H3ABioNet and H3Africa has also presented opportunities for peer support and collaborative research and promoted scientific collaboration between African researchers.

REFERENCES

1. Muir P, Li S, Lou S, et al. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol 2016;17:53.
2. The H3Africa Consortium. Enabling African scientists to engage fully in the genomic revolution. Science 2014;344:1346–8.
3. Mulder NJ, Adebiyi E, Alami R, et al. H3ABioNet, a sustainable pan African bioinformatics network for human heredity and health in Africa. Genome Res 2015;26:271–7.
4. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013;14:178–92.
5. Foster I. Globus Online: accelerating and democratizing science through cloud-based services. IEEE Internet Comput 2011;15:70–3.
6. Lappalainen I, Almeida-King J, Kumanduri V, et al. The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet 2015;47:692–5.
7. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42:377–81.
8. Mulder N, Nembaware V, Adekile A, Inusa B, Brown B, Campbell A. Proceedings of a Sickle Cell Disease Ontology Workshop—towards the first comprehensive Ontology for Sickle Cell Disease. Appll Trans Gen 2016;9:23–9.
9. Horridge M, Tudorache T, Nuylas C, Vendetti J, Noy NF, Musen MA. WebProtégé: a collaborative Web-based platform for editing biomedical ontologies. Bioinformatics 2014;30:2384–5.
10. UniProt Consortium UniProt: a hub for protein information. Nucleic Acids Res 2015;43:D204–12.
11. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucl Acids Res 2001;29:308–11.
12. Hubbard T, Barker D, Birney E, et al. The Ensembl genome database project. Nucl Acids Res 2002;30:38–41.
13. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucl Acids Res 2000;28:235–42.
14. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucl Acids Res 2015;43:D1079–85.
15. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucl Acids Res 2014;42:D980–5.
16. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). Hum Mut 2000;15:57–61.
17. Afgan E, Baker D, van den Beek M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 2016;44:W3–10.
18. Brown DK, Penkler DL, Musyoka TM, Tastan Bishop Ö. JMS: An Open Source Workflow Management System and Web-Based Cluster Front-End for High Performance Computing. PLoS One 2015;10:e0134273.
19. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 2009;5:e1000529.
20. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods 2012;9:179–81.
21. Chimusa ER, Daya M, Möller M, et al. Determining ancestry proportions in complex admixture scenarios in South Africa using a novel Proxy Ancestry Selection Method. PLoS One 2013;8:e73971.
22. Chimusa ER, Mbiyavanga M, Mazandu GK, Mulder NJ. ancGWAS: a post genome-wide association study method for interaction, pathway, and ancestry analysis in homogeneous and admixed populations. Bioinformatics 2016;32:549–56.
23. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 2012;7:e46688.
24. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Gen 2013;76:7.20.1–7.20.41.
25. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinform 2006;22:2729–34.
26. Shihab HA, Gough J, Cooper DN, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using Hidden Markov Models. Hum Mut 2013;34:57–65.
27. Masso M, Vaisman II. AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. Prot Eng Des Sel 2010;23:683–7.
28. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 2006;62:1125–32.
29. Hatherley R, Brown DK, Glenister M, Tastan Bishop Ö. PRIMO: An Interactive Homology Modeling Pipeline. PLoS One 2016;11:e0166698.
30. Campbell MC, Tishkoff SA. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet 2008;9:403–33.
31. Chimusa ER, Meintjies A, Tchanga M, et al. A genomic portrait of haplotype diversity and signatures of selection in indigenous southern African populations. PLoS Genet 2015;11:e1005052.
32. Tastan Bishop ÖT, Adebiyi EF, Alzohairy AM, et al. Bioinformatics education—perspectives and challenges out of Africa. Brief Bioinf 2014;16:355–64.