**TRAINING PLAN**

**Bioinformatics core facility**

**African Centre of Excellence for Genomics of Infectious Diseases (ACEGID)**

1. **Background**

The use of next-generation sequencing (NGS) technologies has now shifted from proof-of-concept studies to adoption as the optimum approach in a variety of molecular biology fields including infectious disease diagnostics, investigation of origins of outbreaks, monitoring transmission, routine pathogen genomic surveillance, and study of pathogen evolution and adaptation. Pathogen genomic surveillance can provide information for timely and effective public health interventions. Prior to and at the onset of COVID-19 pandemic (caused by SARS-CoV-2), pathogen genomics was not well-established in many countries in Africa; it was only an active area of research, and a few institutions and countries were prepared to rapidly implement and adopt large-scale SARS-CoV-2 genomic surveillance. There is compelling evidence for this from the Africa CDC – Africa Pathogen Genomics Initiative (Africa PGI). There still exist scarcity of skilled workforce, high per-sample cost of NGS, and limited number of sequencers. In additionally, most (70%) of the available sequencers existed outside the national public health institutions (NPHIs) and national reference laboratories (NRLs), thereby limiting the application of NGS technologies at the NPHIs and NRLs for infectious disease detection, surveillance.

The timely and accurate detection of the infectious pathogens (viruses, bacteria etc.) is a priority which can only be made possible by using genomic surveillance in management of diseases in our population. The COVID-19 pandemic has highlighted the need for a new public health order to enhance public health surveillance, outbreak detection, and response in Africa – including ACEGID’s leadership to identify priorities and develop strategies to build and strengthen public health workforce and to confront infectious disease threats. Africa continues to grapple with frequent rates of antimicrobial resistance, non-communicable diseases, emerging and re-emerging infections, and endemic diseases every year. While genomics-based surveillance could be utilized to control and eliminate these diseases, genomics capacity for many public health programs remains low in many African countries. There is an urgent need to maximize the benefit of genomics in the continent, and ACEGID jointly with partners began supporting various NGS training programs to serve as both in-country and regional reference site in West Africa for pathogen sequencing.

The ACEGID bioinformatics core facility main undertakings will be to offer service and bioinformatics trainings on the topics such as laboratory information management system, sample management and referral, malaria genomic epidemiology and data analytics, bioinformatics for pathogen genomic surveillance, SARS-CoV-2 genome sequencing and bioinformatics analysis, and new NGS protocols and workflows from the Oxford Nanopore Technologies (ONT), PacBio, and Illumina etc. The core facility will expose the trainees to state-of-art bioinformatic data analysis skills, equipping them with best dry laboratory practice, and providing them with knowledge that they can cascade after their graduation hence ensuring continuity of high-level scientific understating of health and disease in Africa.

Our capacity building plan will focus on bioinformatic training which include use of the Unix command-line and bioinformatics resources (tools and workflows) for sequence data analysis and presentation of data to inform policy built on skillful knowledge. The bioinformatics topics will range from raw read data quality check and variant analyses to data uploading on publicly available databases such as the Global Initiative on Sharing All Influenza Data (GISAID) and the European Nucleotide Archive (ENA) platforms. Overall, the training workshops aims at equipping the individuals with the essential bioinformatics expertise needed to analyze NGS data for implementation of a genomics-informed, real-time, global pathogen surveillance system.

1. **Bioinformatic training overview**

The purpose of these trainings will be to provide intensive hands-on bioinformatics trainings ranging from basic introduction to bioinformatics to advance use of complex bioinformatic analysis tools and workflow management. The trainee will gain the capabilities to use using the Unix operating system, install and configure bioinformatics software, independently handle, analyze, interpret, visualize, and share genomic sequence data on public repositories and using computational resources. The course will be organized in 4 core modules, namely:

**Module 1: Background information** – Thiswill focus on providing an overview on basic terminologies used in bioinformatics and both sequencing technologies, with the goal of covering end-to-end NGS workflows for analyzing genomic data.

**Module 2: Unix/Linux command-line and bash shell** – Thiswill include individual and group hands-on training exercises and practical sessions on Unix/Linux command-line and bash shell (including automation of jobs).

**Module 3: Cluster computing** –Installation and validation of packages for analysis and performing cluster maintenance. How to effectively use the HPC environment to analyze big genomic data.

**Module 4: Bioinformatics and computational analysis and applications** – This entail NGS data processing and workflows (including use of Galaxy) specific to technology (PacBio, ONT, illumine, Hi-C etc). Thiswill include quality control (QC), genome alignment, genome assembly (*De novo* or reference guided), and phylogenetic analysis of the sequenced samples – using workbench tools, Galaxy, web-based tools such as the Nextclade and Phylogenetic Assignment of Named Global Outbreak Lineages (PANGOLIN), and computational workflows. Moreover, participants will learn to share the data through open-access databases such as GISAID and ENA. The bioinformatics part of these training will include lectures and hands-on practical dry-lab sessions. Participants will gain theoretical and practical experience in bioinformatics analysis, interpretation, and sharing of pathogen sequencing data.

**Module 5: Beyond the dry lab –** The participants will learn how to disseminate genomic surveillance findings to policy makers and the public. How to communicate science to non-technical people.

1. **Objectives and expected outcomes.**

The purpose of this course is to provide intensive hands-on training for ACEGID students and staff/researchers on big data analyses and reporting. Participants will gain the capabilities to independently analyze clinical and biological data. Genomic data sharing to promote open science will also be emphasized.

At the end of these trainings, the students and the staff should be able to:

1. To gain competency in genomic analysis – including molecular surveillance,
2. To improve and expand competency in ability to handle, analyze, visualize, and interpret NGS data using bioinformatics tools and resources,
3. To train participants on how to use publicly available databases for genome data sharing,
4. To gain insight on application of NGS data to inform outbreak, epidemic, and pandemic response, and
5. To train participants on how to disseminate genomic surveillance findings to policy makers and the public.
6. **Target audience – Profile of participants**

This course is a targeted training that is designed to support the ACEGID laboratory members and the collaborators from African countries and beyond.

1. **What do you require to attend?**

The training will be done remotely and in person. The participants are encouraged to bring a working laptop – with a good RAM (preferably at least 8 GB) and can support a stable internet connection. The training will be funded by ACEGID and its partners.