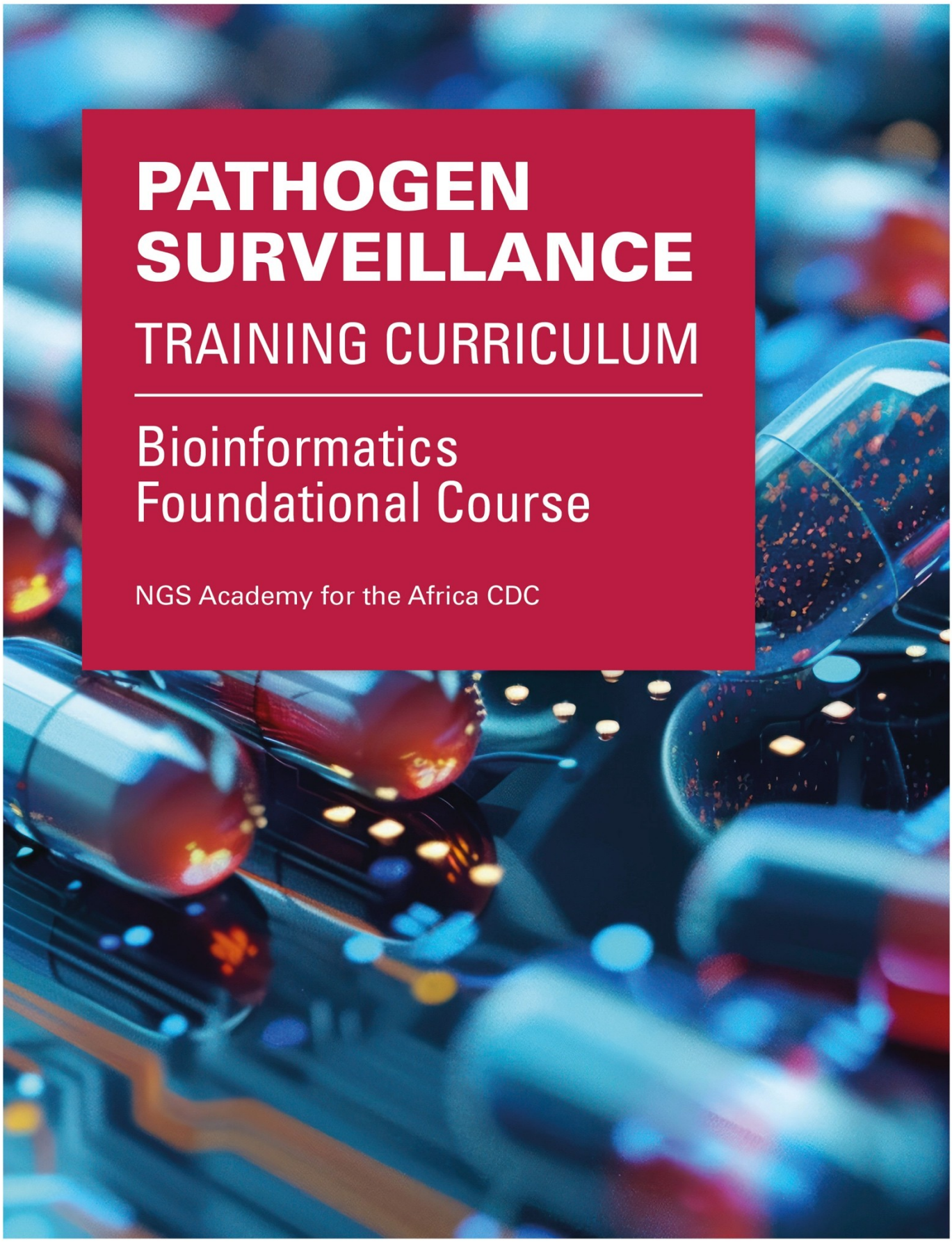
AfricaCDC
Centres for Disease Control and Prevention

Safeguarding Africa's Health



PATHOGEN SURVEILLANCE

TRAINING CURRICULUM

Bioinformatics Foundational Course

NGS Academy for the Africa CDC

ASLM
AFRICAN SOCIETY FOR LABORATORY MEDICINE

H3ABioNet
Pan African Bioinformatics Network for H3Africa

ASLM ACADEMY

Introduction

The Bioinformatics Foundational Course forms part of the larger Pathogen Surveillance Training Curriculum, compiled by the Next-Generation Sequencing Academy for the Africa CDC (NGS Academy for the Africa CDC). Please refer to the Curriculum Overview for a description of the aims, target audiences, and competency level descriptions within this curriculum.

The primary objective of the Bioinformatics Foundational Course is to provide trainees with a comprehensive understanding of various aspects of Bioinformatics, with a specific focus on utilising next-generation sequencing (NGS) technologies for genomic epidemiology.

This course is designed to facilitate the systematic and structured development of knowledge and skills in computational techniques. (See Figure 1) Through a combination of theoretical and practical approaches, the program aims to equip participants with the necessary skills to navigate diverse bioinformatics workflows and utilize relevant tools. The ultimate goal is to build foundational skills to enable trainees to develop and apply computational methods to data processing, analysis and interpretation. The course is aimed at trainees who will spend the majority of their time on data processing and analysis or development/application of bioinformatics tools and workflows for pathogen surveillance.

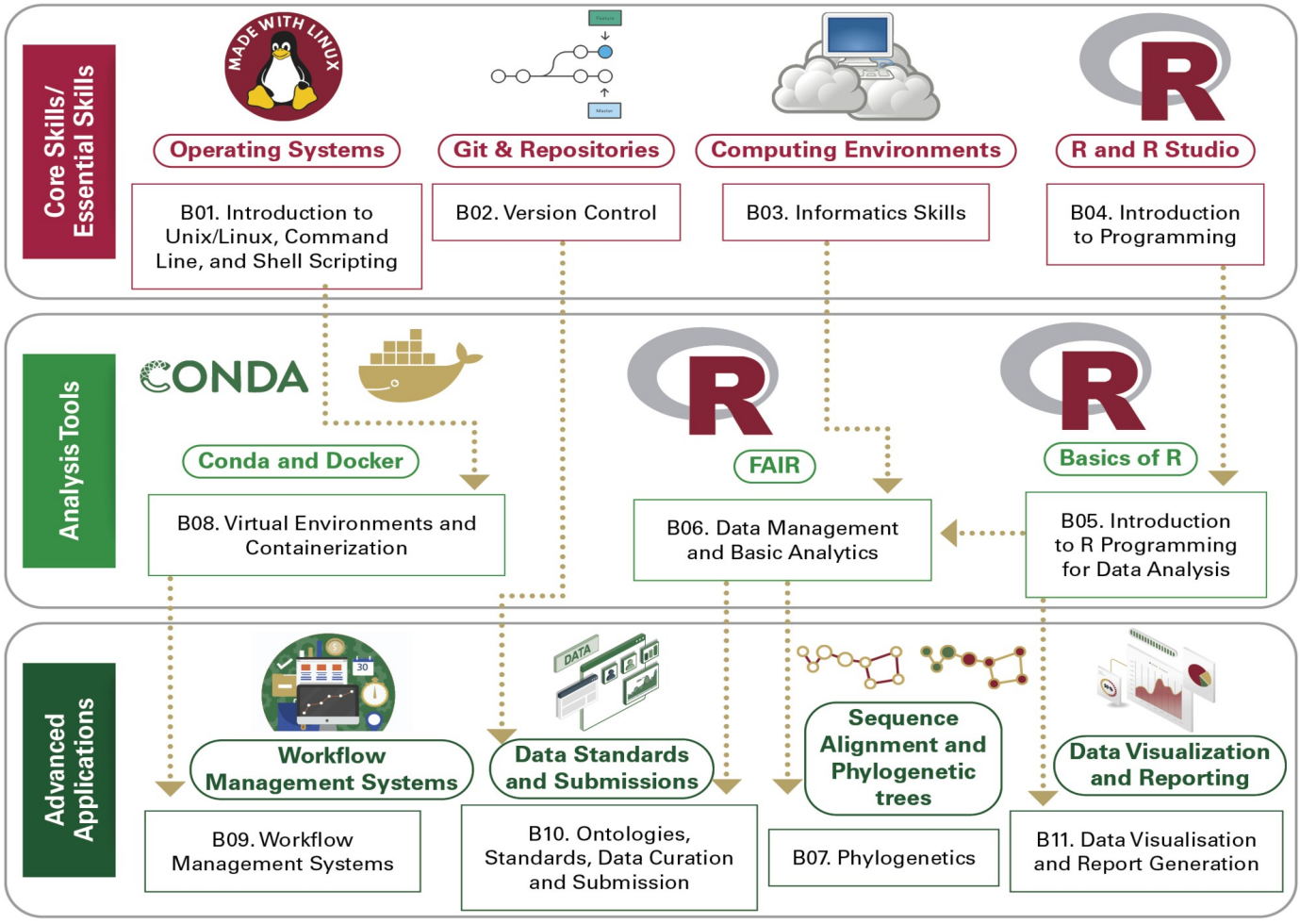


Figure 1. A Learning Pathway for Bioinformatics. Foundational skills (red), analysis tools (green), and advanced applications (dark green) are organised by level, with arrows showing recommended progression.

Scope and content of the course

The course aims to introduce trainees to basic Unix commands and shell scripting so that they can manage files and manipulate data. Some participants will go on to use shell scripting for all their needs, while others may choose to extend their programming skills to Python, R or other programming languages. For those who are responsible for their IT set up and managing workflows, the course includes modules on informatics skills, containerization and workflows. An R module introduces data analytics and is followed by a data management and basic analytics module to cover some basic statistical analysis. Modules have been included to introduce best practices, such as the use of version control, FAIR data management and ontologies and standards. A separate module has been developed to provide the fundamentals of phylogenetics and interpreting phylogenetic trees. More detailed phylogenetics applications to different pathogens are provided in the pathogen-specific courses. A final module teaches trainees how to visualise data for interpretation and reporting. While taking all modules in this course would provide comprehensive foundational bioinformatics skills, trainees may choose to omit some if, for example, they will be using standard workflows and do not need extensive programming skills.

The following table provides the recommended sequence of completion, as well as an overview of the content of the modules within the Bioinformatics Foundational Course.

Module number	Module name	Module overview
B01	Introduction to Unix/Linux, Command Line, and Shell Scripting	This module introduces students to the Unix commands required to manage and manipulate files and directories, and to Shell scripting for basic data analytics.
B02	Introduction to Version Control	This module will familiarize students with code repositories and how they can be used for version control. It covers installing Git, creating a Github Account and directory, as well various aspects of managing code versions.
B03	Informatics Skills	This module aims to provide an overview of computing technology to enable users to identify, set up and troubleshoot computing environments for running bioinformatics software. It covers installing operating systems and software and touches on data security.
B04	Introduction to Programming	This module introduces students to the basics of programming, writing scripts, understanding data structures, types of variables, and using loops. Using programming interfaces and existing libraries and debugging code are included.
B05	Introduction to R Programming for Data Analysis	This module aims to introduce students to the basics of R programming and RStudio for data analysis. It covers installation of R and RStudio, basic programming in R, loading data, and data wrangling using Bioconductor.
B06	Data Management and Basic Analytics	This module described the principles of data management, including data management plans and FAIR data. It also introduces data cleaning and basic statistical concepts for data exploration.
B07	Phylogenetics	This module introduces students to the theory of evolution and phylogenetics, and building multiple sequence alignment and phylogenetic trees.It teaches students how to draw and interpret phylogenetic trees.
B08	Virtual Environments and Containerization	This module provides an introduction to virtualization and containerization using Conda, Bioconda and Docker as examples. It describes how to use and troubleshoot containers effectively, or to containerize tools for analyses.
B09	Workflow Management Systems	This module provides an introduction to workflow management systems, the benefits of workflows and how to develop one. Snakemake and/or Nextflow are used as examples.

Module number	Module name	Module overview
B10	Ontologies, Standards, Data Curation and Submission	This module aims to introduce students to ontologies, standards and tools for data curation and submission. It includes biological/pathogen databases, and how to find and retrieve data from them.
B11	Data Visualisation and Report Generation	This module aims to describe different data visualisation techniques and tools and which to use for different applications. It covers data import, processing, visualization and report creation.

Yellow indicates that the module is aimed at dry laboratory personnel (e.g., epidemiologists, bioinformatics scientists, and bioinformaticians).

Course glossary

D

DNA: Deoxyribonucleic acid

A

Argument: added to a Command to modify the output there is always a space between a command an the argument.

C

Command: the function or script you are trying to run.

Configurability: Containers can be sized to take advantage of more resources (memory, CPU, etc.) on large systems (clusters) or less, depending on the circumstances.

D

Documentation: There is a clear record of what software and software dependencies were used, from bottom to top.

P

Portability: The container can be used on any computer that has Docker installed – it doesn't matter whether the computer is Mac, Windows or Linux-based.

Prompt: The text next to where you type your commands prompts can be modified to include additional information like hostname or current folder location.

R

Reproducibility: You can use the exact same software and environment on your computer and on other resources (like a large-scale computing cluster).

S

Standard error: the error of a failed command

Standard out: the result of a command.

Suggested training methods for this course

- Interactive lectures
- Practical hands-on exercises
- Group discussions
- Analyses using publicly available datasets or specific use cases

References

- ClaudeAI. (2024). ClaudeAI response on Bioinformatics learning path flow diagram. Retrieved Dec 20, 2024, from https://claude.ai.

Acknowledgements

We would like to acknowledge the institutions, listed below, for their financial support which enabled us to design this course. Their generous support made it possible for staff of the NGS Academy for the Africa CDC and Africa CDC – Africa PGI to dedicate the necessary time and resources to further develop this course.

- **Bill & Melinda Gates Foundation** (BMGF, grant numbers: INV-018278, INV-018978, INV-033857, INV-047157 and INV-018977)
- **European Union Health Emergency Preparedness and Response Authority** (EU HERA; grant number: 101145945)
- **Foundation for Innovative Diagnostics** (FIND; grant number: 4419-23-1-FIND-PG)
- **Public Health Alliance for Genomic Epidemiology** (PHA4GE, grant number: 2022-316484)

Last updated: December 2024