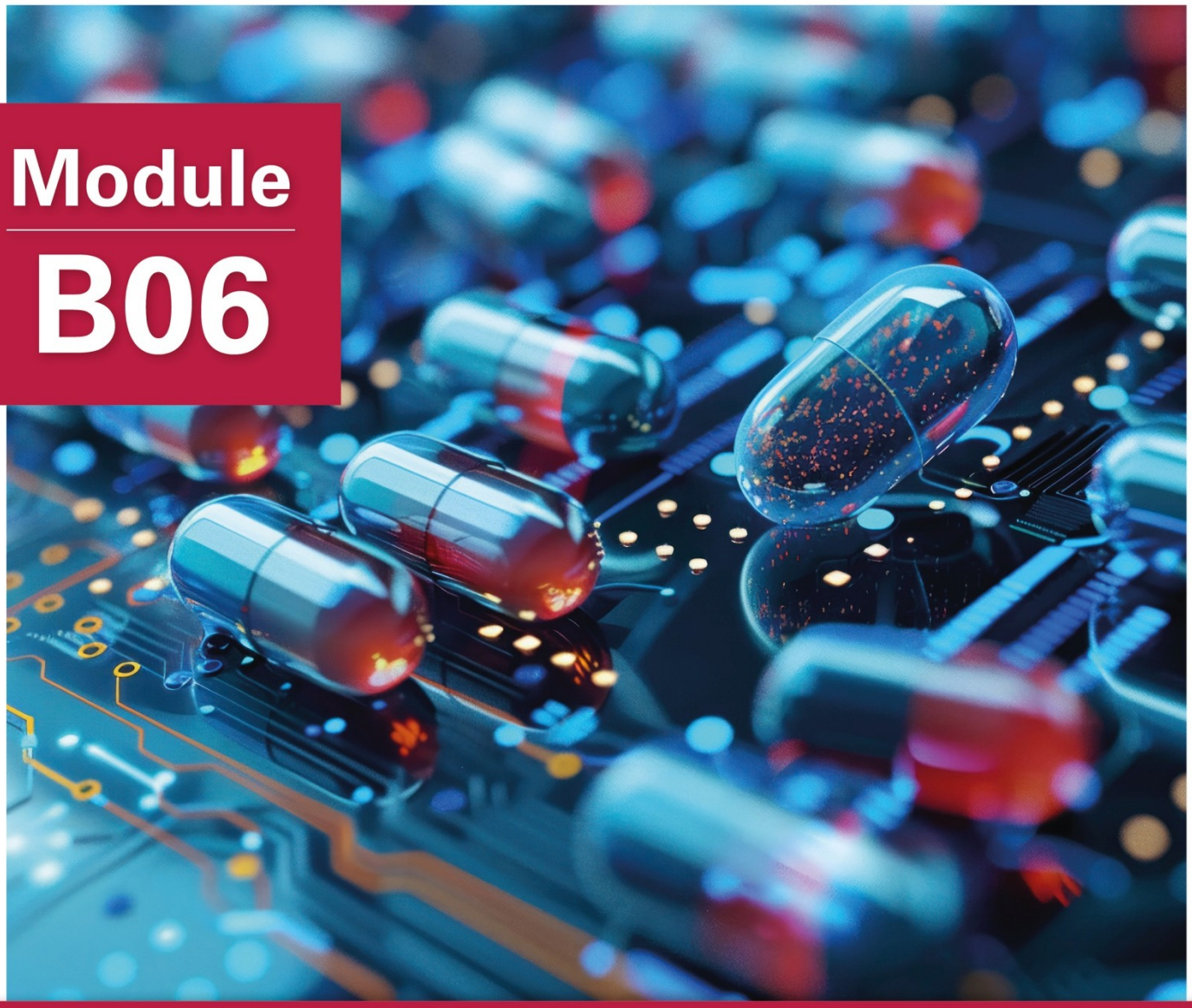
AfricaCDC
Centres for Disease Control and Prevention

Safeguarding Africa's Health



Module B06

Bioinformatics Foundational Course

Data Management and Basic Analytics

NGS Academy for the Africa CDC

ASLM
AFRICAN SOCIETY FOR LABORATORY MEDICINE

H3ABioNet
Pan African Bioinformatics Network for H3Africa

ASLM ACADEMY

Module B06

Data Management and Basic Analytics

back to the table of modules

Module last updated:
December 2024

Suggested or approximate number of sessions	3
Suggested or approximate total learning time	8 hours
Target audience	Bioinformaticians
Delivery format	Lectures, videos, practicals , group discussions
Level of the module	Introductory

Contributors

George Githinji, Tony Yiqun Li, Perceval Maturure, Kennedy Mwai, Nicola Mulder and Sumir Panji.

Module description

This module covers topics in general data management, as well as preparing data for analysis, and basic statistics. Data management is critical to ensure data quality, provenance and accessibility. The module covers creation of data management plans and the principles of FAIR (findable, accessible, interoperable and reusable). Data management in terms of preparation for analysis is also essential, acknowledging that data often undergoes transformations from its initial form to the version ultimately used for primary analyses. A fundamental aspect of data analysis is determining the specific question or objective to be addressed with the data. Approaching the analysis with a hypothesis-driven mindset is crucial, as experiments designed with a clear direction tend to yield more insightful results and are easier to plan compared to those without a specific goal. A key aspect of the data workflow involves data preparation, which includes identifying and correcting errors and ensuring consistent formatting. This step requires meticulous attention to reproducibility, similar to the analysis itself. It is important to consider the importance of saving both raw and intermediate forms, documenting all steps for data provenance, and creating tidy data amenable to analysis.

Recommendations for effective data management/preparation revolve around two main themes. The first theme is to gradually progress towards ready-to-analyze data, taking incremental steps and diligently documenting both the intermediate data and the process involved. By following this approach, you can ensure traceability and reproducibility throughout the data analysis pipeline. The second theme revolves around the concept of "tidy data," which refers to a structured and standardized format for data representation. Tidy data exhibits key features that streamline and expedite the analysis process. By adhering to the principles of tidy data, you can enhance the efficiency and effectiveness of your data analysis endeavors.

Following data preparation and management, the module delves into basic data analytics, covering topics such as file and data formats, tool selection based on format, data cleaning (using OpenRefine as an example), and basic statistics. These can be demonstrated using different tools, including R, Python or Julia (depending on trainer preference or trainee needs). In this module, participants are also introduced to the following topics and/or concepts:

- Basics of data management: data management plans, curating metadata, and data provenance
- FAIR principles for data, what does FAIR mean in practice?
- Data and file formats
- Data cleaning using OpenRefine
- Using OpenRefine to generate and export JSON code for work done in an analysis session
- Importing JSON code file to apply the analysis to another dataset
- Saving an OpenRefine project as a shareable file
- Undoing and redoing actions and exporting the history of actions
- Saving cleaned data in a widely supported file format
- Reproducibility, analysis, question types, the central dogma of inference
- Basic statistics: testing, prediction, variation, experimental design, confounding, power, sample size
- Exploratory data analysis, different types of statistical testing, correlation, regression, causation, and degrees of freedom.
- Using inferential statistics on a single variable approximating a normal distribution

Module learning outcomes

On completion of this module, the participants will have a basic knowledge of, or will be able to:

- Describe the FAIR principles as they relate to data
- Plan their own data management strategy
- Prepare data for analysis starting from data acquisition and focusing on quality control measures
- Clean and explore the data produced for pathogen surveillance
- Define types of variables, data structures, sampling, and collection to inform the statistics to be used
- Manage projects in OpenRefine, including create, export, and import projects
- Work with subsets of data and perform data cleaning and transformation tasks
- Describe concepts such as: reproducibility, analysis, statistics, question types, the central dogma of inference, prediction, variation, experimental design, confounding, power, sample size, correlation, causation, and degrees of freedom.
- Use inferential statistics on a single variable approximating a normal distribution
- Understand linearity, covariance and regression between 2 variables

Module assessments

Module practical: Practical available on the ASLM platform
Module quiz: Assessment questions available on the ASLM platform

Module resources

- The Carpentry | Webpage - Data Carpentry Lessons
- The Carpentry | Webpage - Filtering and Sorting with OpenRefine
- The Carpentry | Webpage - Using scripts
- The Carpentry | Webpage - Exporting and Saving Data from OpenRefine
- The Carpentry | Webpage - Variant Calling Workflow
- The Carpentry | Webpage - Trimming and Filtering
- The Carpentry | Webpage - Assessing Read Quality
- The Carpentry | Webpage - Background and Metadata

Acknowledgements

We would like to thank the following individuals, in alphabetical order of last name, for their valuable time and effort spent in designing (i.e., drafting, reviewing, and refining) this module: **George Githinji, Tony Yiqun Li, Perceval Maturure, Kennedy Mwai, Nicola Mulder and Sumir Panji**.

Furthermore, we would like to thank the following institutions, societies, journals and individuals from whom we sourced open-access resources, used in this module:
The Carpentry.