# **Final Group Project Assignment: Curation and Analysis of Rift Valley Fever Virus (RVFV) Genomic Metadata from Africa**

This project simulates a real-world bioinformatics workflow for genomic metadata extraction, curation, and analysis using GenBank files of Rift Valley Fever Virus (RVFV) isolated from African countries. Each group will retrieve RVFV GenBank records using provided accession numbers, extract structured metadata, clean and analyze the data using Bash, Python, and R, and produce a reproducible report using either Quarto, RMarkdown/R notebook or Jupyter Notebook.

**Analysis time (intermediate level): 3 hours (180 min)**

### **Objectives**

* Demonstrate mastery of Linux/Bash for file manipulation
* Use Python (Biopython) for GenBank parsing and metadata extraction
* Apply R (tidyverse) for data cleaning and visualization
* Practice reproducible reporting with Quarto, JupyterLab or RMarkdown
* Document the complete pipeline and Conda environment applying QMS principle throughout the project

### **Dataset Provided**

* **rvf\_africa.tsv**: Tab separated file with GenBank accession numbers of RVFV isolates from African countries.

### **Required Tools**

* Linux / Bash (e.g., awk, cut, grep, csvkit)
* Python with Biopython, Pandas
* R with tidyverse (dplyr, tidyr, ggplot2)
* Conda for environment management
* Quarto, jupyterlab or RMarkdown for reporting

### **Tasks**

#### **1. Environment Setup**

* Use the Conda environment with required tools.
* Export your \*.yml file

#### **2. Metadata Preparation with Bash**

#### Organize project directories and output data using linux command line

* Inspect and understand the file content using appropriate linux utilities
* How many unique Genbank Accession numbers does the file contain?
* How many sequences are complete and how many are partial?
* Report the number of S, L and M segment in that file
* How many segments of type S, L or M are complete sequences?

**3. GenBank Download (Python + Biopython)**

* Write a script to fetch and download all genbank files for RVF isolated in Africa including all viral segments (see provided metadata file)
* Fetch only complete genomes (e.g. no genbank file for partial segment)
* Extract metadata fields: Organism, Strain, Segment, Host, Country, Collection Date, Product, Protein ID.
* Output clean metadata in a TSV format
* Download fasta files of all L segments. What is the average length of L segment sequences?

#### **4. Data Cleaning, wrangling & Analysis (R tidyverse)**

* Remove entries missing key fields (e.g., no collection date or country) an save filtered data into your working directory
* Group and summarise number of sequences by country, year, and segments
* Mutate to create derived columns (e.g., region from country).
* Count number of sequences by host\_common name
* Filter records by specific countries, years, and segments.
* Rename collection\_year to year
* Group and summarise by country, year, and segment.
* Which countries have the most counts of complete sequences irrespective of the segment
* Count top occurring hosts.

**Perform Descriptive analysis:**

* Number of isolates by country
* Number of isolates per year
* Segments (S, M, L) distribution by country

#### **5. data Visualization**

Integrate at least two boxplot figures into your analysis report:

* Distribution of collection dates by country (boxplot)
* Distribution of collection years by segment
* Summarize number of isolates per country/year
* Plot barplots or maps of strain/country distribution

**Plots must be clearly labeled with legends, titles, and themes.**

#### **6. Reproducible Report (Quarto/RMarkdown/Jupyter)**

* Integrate scripts, outputs, and interpretations in one document.
* Include tool versions, commands, and readable code blocks.

#### **7. Documentation & Packaging**

* README.md describing:  
  + Project goal
  + Tools and versions used
  + Folder structure and how to reproduce

**8. Group Presentation (15 minutes)** Each group will present their pipeline, key results, and quality considerations.

**Format:** Slides (PDF/PPT) or a shared live walkthrough of your notebook or Quarto document.

**9. Evaluation Criteria:**

**Each group will be evaluated based on:**

* **Correctness of analysis steps**
* **Application of best practices quality standards**
* **Completeness and organization of project files**
* **Clarity and accuracy of interpretation**
* **Documentation and presentation quality**

### **Notes**

* All tasks must be performed within the Conda environment.
* You can use the provided python scripts to fetch genbank and to extract metadata
* Reports must be clearly written and self-contained.
* Credit will be given for creativity, clarity, and completeness.
* Best practices, QMS and reproducibility standards will be credited
* The dplyr and boxplot tasks are required components of your analysis section and will be evaluated accordingly.