# **Projet de groupe final : Conservation et analyse des métadonnées génomiques du virus de la fièvre de la vallée du Rift (VFVR) en Afrique**

Ce projet simule un flux de travail bioinformatique réel pour l'extraction, la conservation et l'analyse de métadonnées génomiques à l'aide de fichiers GenBank du virus de la fièvre de la vallée du Rift (VFVR) isolé dans des pays africains. Chaque groupe récupérera les enregistrements GenBank du VFVR à l'aide des numéros d'accès fournis, extraira les métadonnées structurées, nettoiera et analysera les données avec Bash, Python et R, et produira un rapport reproductible via Quarto, RMarkdown/R Notebook ou Jupyter Notebook.

**Temps d'analyse (niveau intermédiaire) : 3 heures (180 min)**

### **Objectifs**

* Démontrer la maîtrise de Linux/Bash pour la manipulation de fichiers
* Utiliser Python (Biopython) pour l'analyse et l'extraction des métadonnées de GenBank
* Appliquer R (tidyverse) pour le nettoyage et la visualisation des données
* Pratiquez le reporting reproductible avec Quarto, JupyterLab ou RMarkdown
* Documenter l'ensemble du pipeline et l'environnement Conda en appliquant le principe QMS tout au long du projet

### **Ensemble de données fourni**

* **rvf\_africa.tsv** : fichier séparé par des tabulations avec les numéros d'accès GenBank des isolats RVFV provenant de pays africains.

### **Outils requis**

* Linux / Bash (par exemple, awk, cut, grep, csvkit)
* Python avec Biopython, Pandas
* R avec tidyverse (dplyr, tidyr, ggplot2)
* Conda pour la gestion de l'environnement
* Quarto, jupyterlab ou RMarkdown pour le reporting

### **Tâches**

#### **1. Configuration de l'environnement**

* Utilisez l'environnement Conda avec les outils requis.
* Exportez votre fichier \*.yml

#### **2. Préparation des métadonnées avec Bash**

#### Organiser les répertoires de projet et les données de sortie à l'aide de la ligne de commande Linux,

* Inspecter et comprendre le contenu du fichier à l'aide d'utilitaires Linux appropriés
* Combien de numéros d'accès Genbank uniques le fichier contient-il ?
* Combien de séquences sont complètes et combien sont partielles ?
* Indiquez le nombre de segments S, L et M dans ce fichier
* Combien de segments de type S, L ou M sont des séquences complètes ?

**3. Téléchargement de GenBank (Python + Biopython)**

* Écrivez un script pour récupérer et télécharger tous les fichiers Genbank pour la RVF isolée en Afrique, y compris tous les segments viraux (voir le fichier de métadonnées fourni)
* Récupérer uniquement les génomes complets (par exemple, aucun fichier Genbank pour un segment partiel)
* Extraire les champs de métadonnées : organisme, souche, segment, hôte, pays, date de collecte, produit, ID de protéine.
* Générer des métadonnées propres au format TSV
* Téléchargez les fichiers fasta de tous les segments L. Quelle est la longueur moyenne des séquences de segments L ?

#### **4. Nettoyage, gestion et analyse des données (R tidyverse)**

* Supprimez les entrées manquantes dans les champs clés (par exemple, aucune date de collecte ou pays) et enregistrez les données filtrées dans votre répertoire de travail
* Regrouper et résumer le nombre de séquences par pays, année et segments
* Muter pour créer des colonnes dérivées (par exemple, région à partir du pays).
* Compter le nombre de séquences par nom commun d'hôte
* Filtrer les enregistrements par pays, années et segments spécifiques.
* Renommer collection\_year en year
* Regrouper et résumer par pays, année et segment.
* Quels pays ont le plus grand nombre de séquences complètes, quel que soit le segment
* Compter les hôtes les plus fréquents.

**Effectuer une analyse descriptive :**

* Nombre d'isolats par pays
* Nombre d'isolats par an
* Répartition des segments (S, M, L) par pays

#### **5. Visualisation des données**

Intégrez au moins deux figures de boxplot dans votre rapport d’analyse :

* Répartition des dates de collecte par pays (boxplot)
* Répartition des années de collecte par segment
* Résumer le nombre d'isolats par pays/an
* Tracer des graphiques à barres ou des cartes de la répartition des souches/pays

**Les intrigues doivent être clairement étiquetées avec des légendes, des titres et des thèmes.**

#### **6. Rapport reproductible (Quarto/RMarkdown/Jupyter)**

* Intégrez des scripts, des sorties et des interprétations dans un seul document.
* Inclure les versions d’outils, les commandes et les blocs de code lisibles.

#### **7. Documentation et emballage**

* README.md décrivant :  
  + Objectif du projet
  + Outils et versions utilisés
  + Structure des dossiers et comment les reproduire

**8. Présentation de groupe (15 minutes)**Chaque groupe présentera son pipeline, ses résultats clés et ses considérations de qualité.

**Format :** Diapositives (PDF/PPT) ou une présentation en direct partagée de votre bloc-notes ou document Quarto.

**9. Critères d’évaluation :**

**Chaque groupe sera évalué en fonction de :**

* **Exactitude des étapes d'analyse**
* **Application des meilleures pratiques en matière de normes de qualité**
* **Exhaustivité et organisation des dossiers de projet**
* **Clarté et précision de l'interprétation**
* **Qualité de la documentation et de la présentation**

### **Remarques**

* Toutes les tâches doivent être effectuées dans l’environnement Conda.
* Vous pouvez utiliser les scripts Python fournis pour récupérer Genbank et extraire les métadonnées
* Les rapports doivent être clairement rédigés et autonomes.
* Le crédit sera accordé à la créativité, à la clarté et à l’exhaustivité.
* Les meilleures pratiques, le SMQ et les normes de reproductibilité seront crédités
* Les tâches dplyr et boxplot sont des composants obligatoires de votre section d'analyse et seront évaluées en conséquence.