Reproducible workflow proposal

Title: Identifying Disease-Causing taxa and Genes Using 16S rDNA Metagenomics Data

What data are you using, and what is its source?

This project will use publicly available 16S rDNA sequencing data from NCBI SRA that are from environmental microbiome studies (e.g., hospital infections, wastewater, and soil).

Describe the data in terms of volume (how big is your data?)

Hundreds of samples (~2–5 GB per sample) with each sample contains millions of 16S rDNA reads. Each sample will contain **millions of 16S rDNA reads**, requiring computational efficiency for taxonomic classification and functional prediction

Basic research question?

In this work will intend to know which microbial taxa are associated with disease-causing genes? and how do microbiome structures differ between healthy and disease-associated samples?

How do I plan to analyze the data?

1. Genome sequences will be analyzed using Bioinformatic pipeline in HPC and R Software
2. Microbiome Composition will be analyzed using these Packages in R: `phyloseq`, `vegan`, `microbiome`, `qiime2R`
    - **Methods**:
        - Alpha diversity (Shannon, Simpson)
        - Beta diversity (Bray-Curtis)
        - Differential abundance analysis (DESeq2)
3. Functional prediction of microbial genes from 16S data
    - **Tool**: `PICRUSt2` to infer functional genes from 16S data.
    - **Databases**: KEGG Orthologs (KO), Virulence Factor Database (VFDB).

How do you plan to turn your data into a reproducible workflow?

- R Markdown for documentation.
- GitHub for version control.
- Upload Bioinformatics pipeline on Github

This study will help identify disease-causing microbes using only 16S rDNA data, enabling better predictions of microbial virulence taxa in environmental samples.