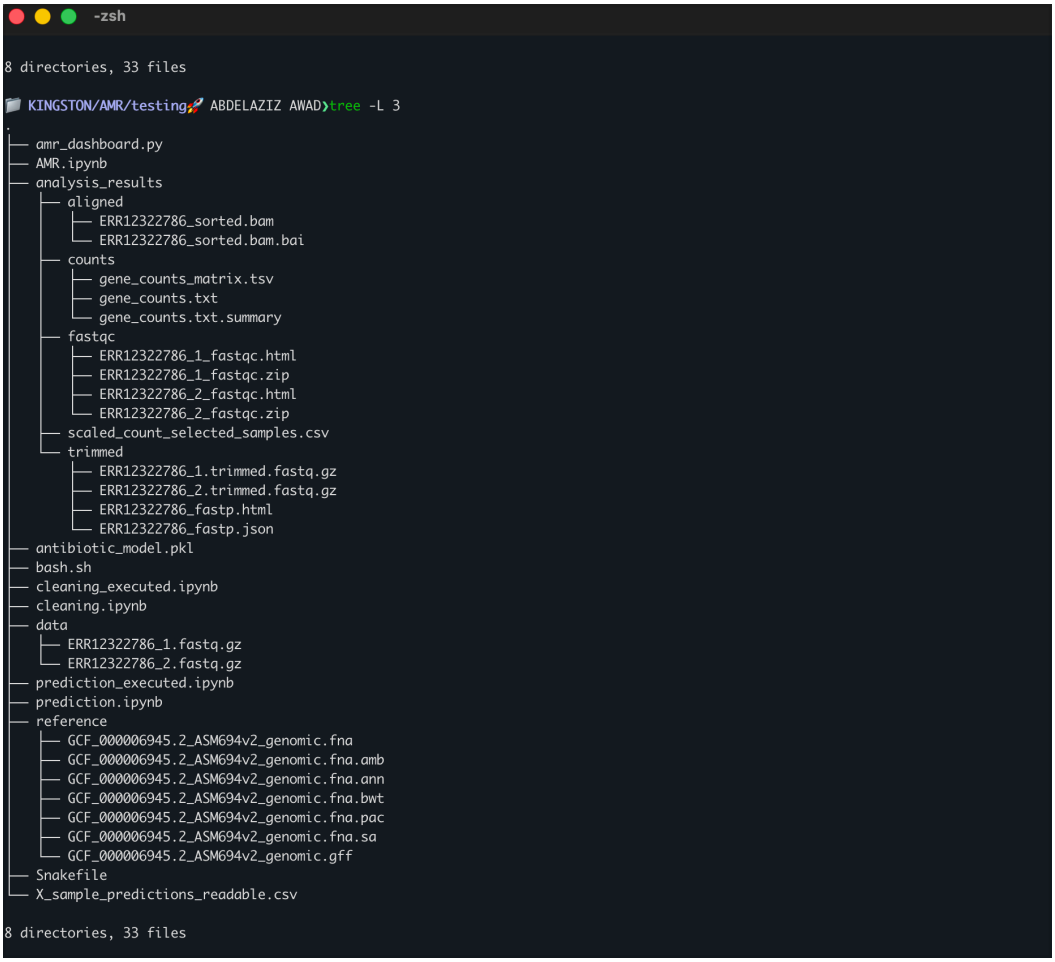
Steps of the project

1. Quality control
2. Trimming
3. Alignment to the reference genome
4. Feature counting
5. Data Engineering
6. Prediction modeling using a multi-output random forest classifier
7. Automating the data extraction process and data engineering
8. Launching the dashboard for interpretable visualization of the predictions

Assumptions for reproducible work:

1. **You must have the above structure to work properly, unless you can work with bash variables properly**
2. **The bash code will install the missing packages; otherwise, you could download it using brew install fastqc fastp bwa samtools brewsci/bio/subread && conda install -y -c conda-forge -c bioconda pandas numpy scikit-learn joblib streamlit plotly snakemake papermill matplotlib seaborn llvmlite numba**
3. **Data: you must have paired-end fastq files, and you must have the reference files; REF="reference/ GCF_000006945.2_ASM694v2_genomic.fna", GTF="reference/ GCF_000006945.2_ASM694v2_genomic.gff"**



Steps for reproducible work:

1. Run the snakemake file to get the matrix ready for AI prediction using: snakemake -j 4

2. Use this code to launch the dashboard for visualization of the results: streamlit run amr_dashboard.py

3.Then upload the engineered matrix you get from the previous code (snakemake code) to the dashboard, it should be: "analysis_results/scaled_count_selected_samples.csv"