

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"

```

-zsh

8 directories, 33 files
# KINGSTON/AMR/testing$ ABDELAZIZ AWAD>tree -L 3
.
+-- amr_dashboard.py
+-- AMR.ipynb
+-- analysis_results
|   +-- aligned
|   |   |-- ERR12322786_sorted.bam
|   |   |-- ERR12322786_sorted.bam.bai
|   +-- counts
|   |   |-- gene_counts_matrix.tsv
|   |   |-- gene_counts.txt
|   |   |-- gene_counts.txt.summary
|   +-- fastqc
|   |   |-- ERR12322786_1_fastqc.html
|   |   |-- ERR12322786_1_fastqc.zip
|   |   |-- ERR12322786_2_fastqc.html
|   |   |-- ERR12322786_2_fastqc.zip
|   +-- scaled_count_selected_samples.csv
|   +-- trimmed
|   |   |-- ERR12322786_1_trimmed.fastq.gz
|   |   |-- ERR12322786_2_trimmed.fastq.gz
|   |   |-- ERR12322786_fastp.html
|   |   |-- ERR12322786_fastp.json
+-- antibiotic_model.pkl
+-- bash.sh
+-- cleaning_executed.ipynb
+-- cleaning.ipynb
+-- data
|   +-- ERR12322786_1.fastq.gz
|   +-- ERR12322786_2.fastq.gz
+-- prediction_executed.ipynb
+-- prediction.ipynb
+-- reference
|   +-- GCF_000006945.2_ASM694v2_genomic.fna
|   +-- GCF_000006945.2_ASM694v2_genomic.fna.amb
|   +-- GCF_000006945.2_ASM694v2_genomic.fna.ann
|   +-- GCF_000006945.2_ASM694v2_genomic.fna.bwt
|   +-- GCF_000006945.2_ASM694v2_genomic.fna.pac
|   +-- GCF_000006945.2_ASM694v2_genomic.fna.sa
|   +-- GCF_000006945.2_ASM694v2_genomic.gff
+-- Snakefile
+-- X_sample_predictions_readable.csv

8 directories, 33 files

```

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"