Machine Learning-Driven Pangenome Pipeline for Predicting *E. coli* Drug Resistance

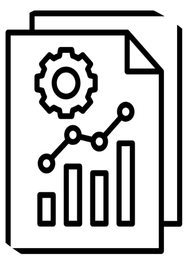
CoDE Lab, San Francisco State University, Departments of Biology & Computer Science

Abdoulfatah Abdillahi aabdillahi@mail.sfsu.edu
Estefanos Kebebew ekebebew@sfsu.edu
Myco Torres mtorres29@sfsu.edu
Juvenal F Barajas jbarajas8@mail.sfsu.edu
Pleuni Pennings, PhD pennings@sfsu.edu

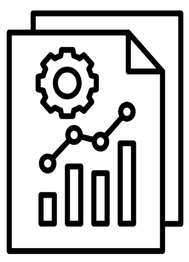
Why We Care About *E. coli*

- **Common and Widespread Pathogen**: Antibiotic resistance in *E. coli i*s a major concern because it is the most common Gram-negative pathogen affecting humans.
- **Increasing Multidrug Resistance**: Rising numbers of E. coli strains are becoming resistant to multiple antibiotics, which significantly limits treatment options and contributes to higher morbidity and mortality rates.
- **Limitations of Traditional Testing**: Current drug resistance tests for E. coli are time-consuming, have low throughput, and are limited to bacteria that can be easily cultivated in labs.
- **Current methods:** Some studies train predictive models using only **gene presence/absence** or **SNP** data. We propose a novel approach that combines both within a pangenome framework to enhance accuracy and improve the generalizability of drug resistance predictions.

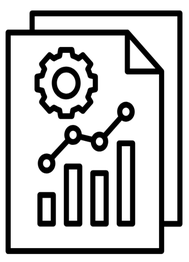
Prepare Pangenome Reference



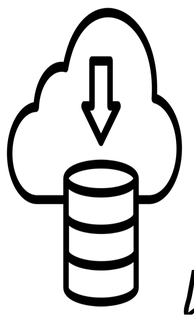
Mills Dataset (2000 Samples)



Gladstone Dataset (3000 Samples)



Kalloneen Dataset (2000 Samples)



Download 200 samples each dataset

Trim FastQ Files

- Trim low-quality reads.
- Filter out adapter

De novo Assembly

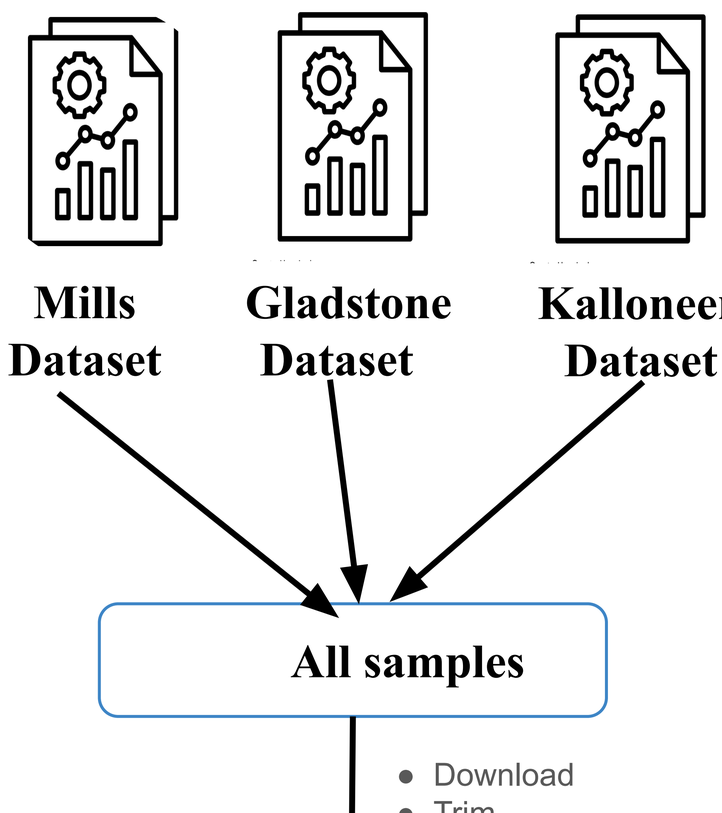
- Combines sequencing data to produce accurate contigs

Annotate Fasta Files

- Annotated genome in GFF file

Pangenome Creation

Pangenome Pipeline / Preprocessing



Feature Extraction

Accession ID	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene …n
ERR4035812	ACGT	GGTA	ACCT	ACTG	TCCA	NNN
ERR4035813	ACG	CTAA	ACCT	ACTG	TCCA	ACTN
ERR4035814	NNN	CTAA	ACCT	ACTG	TACA	ACTN
ERR4035815	TCGT	CTAC	GGT	ACTG	ACTG	NNN
EER…n	ACG	CTAC	GGC	ACTG	NNN	NNN

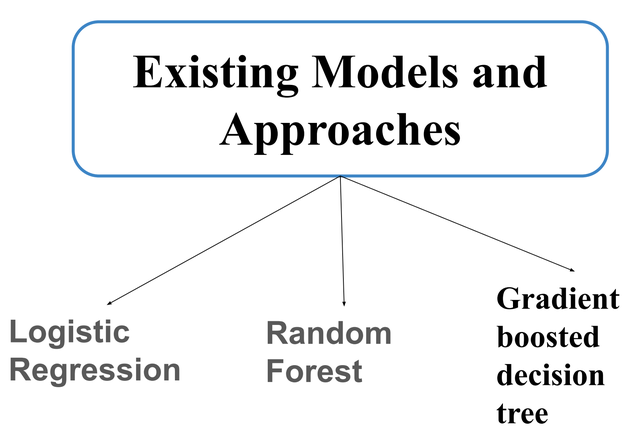
SNP

Gene Presence

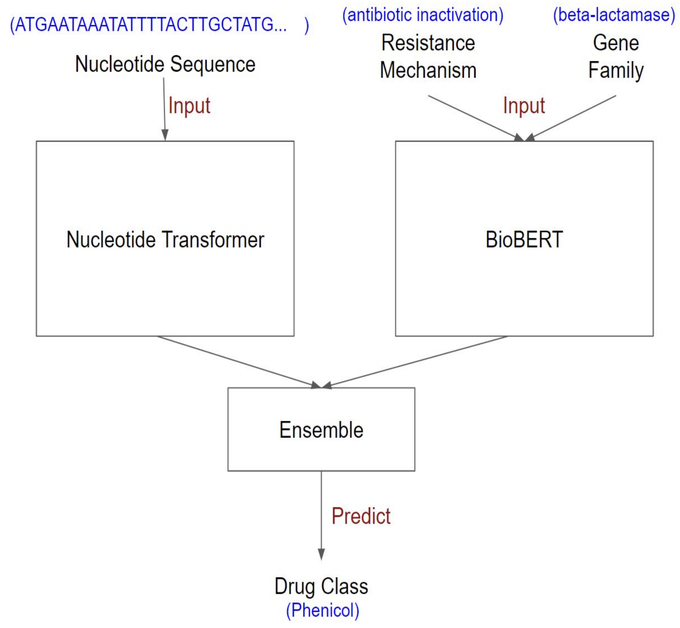
Gene Absence

This table represents our feature extraction process, where we identify SNPs, gene presence, and gene absence across various genes for each accession ID.

Machine Learning



Future Direction using NLP



References

- Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., & Parts, L. (2018). Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. PLoS Computational Biology, 14(12), e1006258. https://doi.org/10.1371/journal.pcbi.1006258
- Yoo, H. (n.d.). *Predicting anti-microbial resistance using large language models*. ar5iv. https://ar5iv.labs.arxiv.org/html/2401.00642

Acknowledgement

Funding:

- Student Enrichment Opportunities
- NIH
- Bristol Myers Squibb
- Kenfong Award