1. Purpose of the database

Create a query system that allows exploration of genetic elements (AMRs, MLSTs, etc.) and metadata in a series of sample collected in different sources (environments).

1. Key Entities

* Samples: Information about each E. coli sample.
* Genes: Details of genes, especially AMR genes.
* Plasmids: Information on plasmids identified in the samples.
* Transposons: Details about transposons.
* Other integrons: Information on integrons.
* Software: Details about the software used for detection.

1. Design Tables

Each entity will correspond to a table in the database.

samples

sample\_id (Primary Key)

source

collection\_date

other\_metadata

genes

gene\_id (Primary Key)

gene\_name

gene\_function (e.g., resistance mechanism)

gene\_sequence

plasmids

plasmid\_id (Primary Key)

plasmid\_name

plasmid\_type

other\_characteristics

transposons

transposon\_id (Primary Key)

transposon\_name

associated\_genes (could be a reference to gene\_id)

integrons

integron\_id (Primary Key)

integron\_type

associated\_elements (this could involve complex relationships with genes, plasmids, etc.)

software

software\_id (Primary Key)

software\_name

software\_version

analysis\_type (e.g., AMR detection, plasmid identification)

1. Relationships

Many-to-Many Relationships: A sample might contain multiple plasmids, and a plasmid might be found in multiple samples. This would require junction tables like sample\_plasmids.

Foreign Keys: For linking related data, such as linking genes detected by specific software.

1. Design Junction Tables for Many-to-Many

Relationships

sample\_plasmids

sample\_id (Foreign Key)

plasmid\_id (Foreign Key)

sample\_genes

sample\_id (Foreign Key)

gene\_id (Foreign Key)

Example to produce a list of AMR genes per sample and per source:

sqlite> SELECT

...> s.sample\_id,

...> s.source,

...> ARRAY\_AGG(g.gene\_name) AS amr\_genes

...> FROM

...> samples s

...> JOIN

...> sample\_genes sg ON s.sample\_id = sg.sample\_id

...> JOIN

...> genes g ON sg.gene\_id = g.gene\_id

...> WHERE

...> g.gene\_function = 'AMR'

...> GROUP BY

...> s.sample\_id,

...> s.source;