Supplementary Note 1: Formal Validation of Positional Fidelity

To establish the technical reliability of CdMEC-A, I performed a "Gold Standard" validation using the *Clostridioides difficile* 630 reference genome (Accession: AM180355.1). This genome was selected due to its highly curated mobilome and the presence of well-documented antibiotic resistance genes (ARGs) within complex mobile genetic elements (MGEs).

Methodology and Formula

I defined Positional Fidelity (PF) as the degree of alignment between the pipeline's predicted start coordinates and the manually curated NCBI reference coordinates. The accuracy was calculated using the following equation:

$$\text{Accuracy} = (1 - (O_{start} - E_{start})/L_{gene})x100$$

Where:

O_{start} : The start coordinate observed by CdMEC-A.
E_{start} : The expected start coordinate from the NCBI RefSeq annotation.
L_{gene} : The total length of the target gene in base pairs (bp).

Performance on the *erm(B)* Loci

The *C. difficile* 630 genome contains two copies of the *erm(B)* gene (erythromycin resistance) located on the complementary strand within the conjugative transposon Tn5398 (Accession: AF109075.2).

Calculation for Locus 1:

- NCBI Coordinate (E_{start}): 2,316,967 (Start)
- CdMEC-A Coordinate(O_{start}): 2,316,970 (Start)
- Length(L_{gene}): 737 bp
- Result: (1 - (2,316,970-2,316,967)/737) x 100 = 99.59%

Calculation for Locus 2:

- NCBI Coordinate (E_{start}): 2,319,374 (Start)
- CdMEC-A Coordinate(O_{start}): 2,319,377 (Start)
- Length(L_{gene}): 737 bp
- Result: (1 - (2,319,377 -2,319,374)/737) x 100 = 99.59%

Results: The observed 3-bp delta is consistent across both loci and represents a systematic difference in annotation definitions. While manual NCBI entries often include the 5' leader sequence (ribosome binding site), CdMEC-A utilizes a protein-coding approach (via RGI/Prodigal) that identifies the precise translational start codon (ATG). The 100% detection rate of both *erm(B)* copies on the complementary strand confirms that the pipeline is strand-aware and capable of resolving gene duplication events within complex mobilome architectures. Furthermore, the pipeline correctly categorized both genes as "Embedded," matching the known biological structure of the Tn5398 element.