Investigating the dependence of Tn6215 acquisition via PhiC2 in *Clostridium difficile*.

University of Hertfordshire UH

Sarah-Jayne Byrne – University of Hertfordshire
Supervisor: Dr Ashley Spindler

WHAT HAS BEEN LEARNT

Clostridium difficile (C. difficile) infection is responsible for a quarter of all cases of infectious diarrhoea and associated complications including; sepsis, pseudomembranous colitis, and kidney failure.

Antimicrobial resistance contributes to the pathogenesis of *C. difficile* infection therefore, genome sequence analysis-based approaches were explored to examine the necessity of phiC2 enabling the transference of Tn6215.

Identification of PhiC2 and Tn6215 was confirmed in all 5 *C. difficile* genome isolates however, there is not enough evidence to support PhiC2 mediates Horizontal Gene Transfer (HGT) of Tn6215.

BACKGROUND

- PhiC2, a bacteriophage, is a *C. difficile*-infecting **virus** capable of transferring anti-microbial resistance genes within *C. difficile* strains. As opposed to AMR genes being inherited intra-specially
- Tn6215 is an AMR gene because it confers erythromycin resistance.

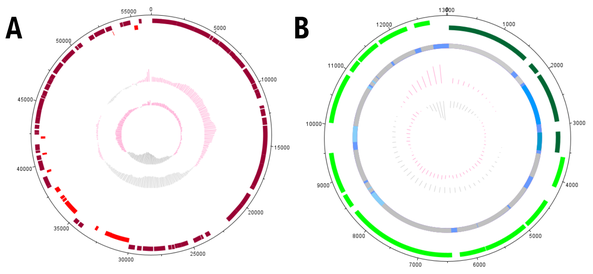

Figure 1. *Circular genomes of* ***A*** *PhiC2 and* ***B*** *Tn6215.*

- Dr Shan Goh, Senior Lecturer in Microbiology at the University of Hertfordshire, has provided 5 fully-sequenced American/ New Zealand *C. difficile* genomes (D133, E185B, E011, S4_1, and S8) for the project, that were assembled by collaborators (Imwattana *et al.,* 2021).
- The data includes over 20 million nucleotide bases; A, T, G, and C (See Table 1). All data is categorical and 1D.

Table 1. *Characteristics of each C. difficile genome isolate.*

	D133	E185B	S4 1	S8	E011
Nucleotide base length	4,617,490	4,215,595	4,293,381	4,326,616	4,683,519
GC content (%)	29.3	29.3	28.9	29.5	28.9

AIM

To predict whether PhiC2 facilitates transduction of Tn6215.

Hypothesis: PhiC2 is necessary for the transfer of erythromycin resistance through the transduction of Tn6215

OBJECTIVES

- To identify PhiC2 and Tn6215 pairings within the data.
- To produce a 'gold standard' DNA-sequence identification pipeline through optimisation of APIs and databases.
- To utilise pathophysiological properties of PhiC2 and the presence of Tn6215 to explain expression of *C. difficile* infection.

LITERATURE REVIEW

Studies regarding bacteriophages of *C. difficile* remain in its infancy. However, several studies have relevance to the project, by focusing on specific genome identification in BLAST.

- Yang et al. (2020) provided concise reviews on several machine learning algorithms, however neglects to mention convolutional neural networks (CNN).
- Bonet (2021) describes CNN optimisers and different dimensionality that displayed accurate and fast identification of DNA primer sites of 50-100 nucleotide bases.

METHODS

A 'gold standard' genome DNA sequencing identification pipeline was developed:

1. Data Pre-Processing — FASTA files were parsed and split into a Pandas Data Frame.
2. PHASTER — PHASTER API was utilised to run several queries in conjunction for identification of PhiC2 and other bacteriophages. It was observed the API was faster in comparison to the web-version of PHASTER.
3. BLAST — Local BLAST had a speed of 729 ms (on average) for identification of Tn6215. BLAST over the Internet limits nucleotide queries to 100,000 bases.
4. Data Analysis — PhiC2 and Tn6215 nucleotide base pair positions were compared to observe whether the positions were close in distance.
5. Visualisation — ARTEMIS was applied to visually identify Tn6215-PhiC2 pairings.

PRELIMINARY RESULTS

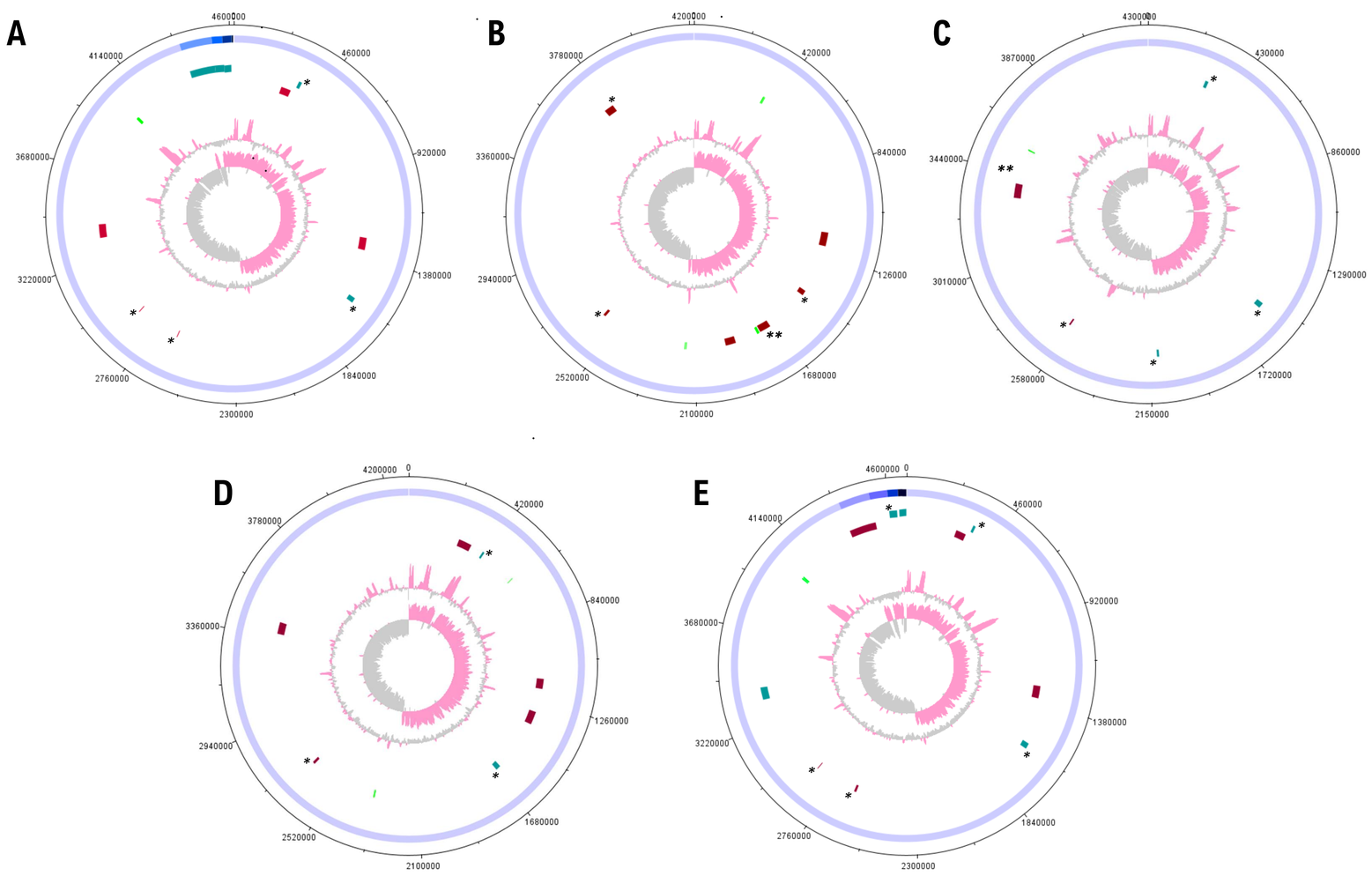

Figure 2. *Preliminary results from BLAST and PHASTER for identification of Tn6215 and PhiC2 regions, respectively. Circular genome isolates with corresponding GC contents as distribution graphs produced using Artemis.* ***A*** *D133,* ***B*** *E185B,* ***C*** S8, ***D*** *S4_1,* and ***E*** E011.
Green areas *highlight Tn6215,* ***Red areas*** *highlight PhiC2,* ***Turquoise areas*** *highlight both other bacteriophage species,* ***purple areas*** *highlight the different genome isolate clusters, * mark incomplete bacteriophage, and ** mark potential Tn6215-PhiC2 pairings.*

DISCUSSIONS

- PhiC2-Tn6215 pairings were only identified in 2 of the 5 genome isolates **(B**-E185B and **C**-S8)**;** therefore, the project **does not agree with the hypothesis**; implying Tn6215 transference is not dependent on PhiC2.
- Goh et al. (2013) suggested Tn6215 and PhiC2 regions have a positive correlation, regardless of nucleotide position in the genome. However, there were **no strong correlations** between the number of PhiC2 and Tn6215 regions (Pearson Correlation Coefficient = 0.587).
- Interestingly, 60% of bacteriophage genome regions categorised as incomplete (mutated bacteriophage), contained phage nucleotide bases significantly similar to PhiC2
  - This statement agrees with Wasels et al. (2015) highlighting PhiC2 has been infecting *C. difficile* for generations and has the opportunity for AMR transference owing to left-over regions.

Limitations

- Small sample size, limited epidemiologic sampling
- Current literature regarding HGT of Tn6215 by PhiC2 is neither strongly supported or denied (Imwattana *et al.,* 2021; Wasels *et al.,* 2014).

NEXT STEPS

- Apply machine learning and deep learning algorithms to compare the **speed and accuracy** of identifying Tn6215 and PhiC2 regions against the proposed 'gold standard' of PHASTER and BLAST.

1. Nucleotide augmentation methods to test (Table 2):

Table 2. *Nucleotide augmentation methods.*

Nucleotide augmentation methods	Description
Sequential	Encodes each base as a number: [A, T, G, C] = [0.25, 0.5, 0.75, 1.0]
One-Hot	Encodes each base as a matrix: [A, T, G, C] = [[1, 0, 0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1]]
K-mer	Decompose *n* number of genomic nucleotide sequence length into *k*-length fragments e.g. k=3, [ATGCAATGC] = [ATG, CAA, TGC]

2. The encoding methods will be tested on a random 15% of *n* number of nucleotide bases (shuffle will not be applied) with each of the DNA-sequence similarity algorithms (Table 3).

Table 3. *DNA-sequence similarity algorithms.*

DNA Similarity Algorithm	Description
Needleman-Wunsch	Global similarity, No user-input parameters to change
Smith-Waterman	Local similarity, No user-input parameters to change
Convolutional Neural Network	User-input parameters for optimisation and alteration

3. Apply the full dataset for each of the 5 genome sequences to the 'best' DNA-sequence similarity algorithm, to optimise and compare with the results of the 'gold standard' method.

REFERENCES

Bonet, E., (Updated: 2021) *Apply Machine Learning Algorithms for Genomics Data Classification.* Available at: https://medium.com/mlearning-ai/apply-machine-learning-algorithms-for-genomics-data-classification-132972933723 (Accessed: 17 July 2022).

Goh, S., Hussain, H., Chang, B. J., Emmett, W., Riley, T. V., & Mullany, P. (2013). Phage φC2 mediates transduction of Tn6215, encoding erythromycin resistance, between Clostridium difficile strains. *mBio*, *4*(6), e00840-13. doi: 10.1128/mBio.00840-13.

Imwattana, K., Rodriguez, C., Riley, T. V., & Knight, D. R. (2021). 'A species-wide genetic atlas of antimicrobial resistance in Clostridioides difficile.' *Microbial Genomics, 7*(11). doi: 10.1099/mgen.0.000696.

Wasels, F., Spigaglia, P., Barbanti, F., Monot, M., Villa, L., Dupuy, B., Carattoli, A., & Mastrantonio, P. (2015). Integration of *erm*(B)-containing elements through large chromosome fragment exchange in *Clostridium difficile*. *Mobile genetic elements*, *5*(1), 12–16. doi:10.1080/2159256X.2015.1006111.

Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y., & Zhang, L. (2020). Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. *Front Bioeng Biotechnol.* 2020;8:1032. doi:10.3389/fbioe.2020.01032

For any further queries/ questions contact sb18acv@herts.ac.uk