Epigenomic characterization of *Clostridioides difficile* finds a conserved DNA methyltransferase that mediates sporulation and pathogenesis

Pedro H. Oliveira[1], John W. Ribis[2], Elizabeth M. Garrett[3], Dominika Trzilova[3], Alex Kim[1], Ognjen Sekulovic[2], Edward A. Mead[1], Theodore Pak[1], Shijia Zhu[1], Gintaras Deikus[1], Marie Touchon[4,5], Martha Lewis-Sandari[1], Colleen Beckford[1], Nathalie E. Zeitouni[1], Deena R. Altman[1,6], Elizabeth Webster[1], Irina Oussenko[1], Supinda Bunyavanich[1], Aneel K. Aggarwal[7], Ali Bashir[1], Gopi Patel[6], Frances Wallach[6], Camille Hamula[6], Shirish Huprikar[6], Eric E. Schadt[1], Robert Sebra[1], Harm van Bakel[1], Andrew Kasarskis[1], Rita Tamayo[3], Aimee Shen[2*], Gang Fang[1*]

Supplementary Notes

Supplementary Notes

Introduction

Clinical symptoms of *C. difficile* infection (CDI) in humans range in severity from mild self-limiting diarrhea to severe, life-threatening inflammatory conditions, such as pseudomembranous colitis or toxic megacolon. Since the vegetative form of *C. difficile* cannot survive in the presence of oxygen, the bacterium is transmitted via the fecal/oral route as a metabolically dormant spore[1]. In the intestinal environment, these spores subsequently germinate into actively growing, toxin-producing vegetative cells that are responsible for disease pathology[2]. CDI progresses in an environment of host microbiota dysbiosis, which disrupts the colonization resistance typically provided by a diverse intestinal microbiota[3]. In the last two decades, there has been a dramatic rise in outbreaks with increased mortality and morbidity due in part to the emergence of epidemic-associated strains with enhanced growth[4,5], toxin production[6], and antibiotic resistance[7]. *C. difficile* was responsible for half a million infections in the United States in 2011, with 29,000 individuals dying within 30 days of the initial diagnosis[8]. Those most at risk are older adults, particularly those who take antibiotics that perturb the normally protective intestinal microbiota.

5mC methylation motifs

While SMRT-seq can effectively detect 6mA and 4mC events, it does not effectively detect 5mC events[9]. Therefore, no confident 5mC motifs were detected from SMRT-seq data. Because MTase gene analysis suggests that <10% of MTases are of the 5mC type across the 36 *C. difficile* genomes, we focused on 6mA and 4mC methylation in this study.

A joint examination of defense systems and gene flux in *C. difficile*

38 MTases (27%) were found to belong to Type I R-M systems, and 100 MTases (72%) to Type II (Fig. 1c). All but one of the Type II MTases are solitary, *i.e.*, devoid of a cognate restriction endonuclease (REase) (Fig. 1d). The least abundant MTases (1%) belong to Type III R-M systems (Fig. 1c). A Type IV (no cognate MTase) McrBC REase gene was also present in all genomes (Supplementary Table 2b). 28% of MTase genes were located in mobile genetic elements (MGEs) (19% in prophages and 9% in integrative conjugative/mobile elements), while the large majority was encoded in other chromosomal regions (Fig. 1e, Supplementary Tables 2b-d). No MTase genes were found in plasmids. While the relationship between some defense systems (e.g., CRISPR-Cas and R-M systems) has been studied in an experimental setting[10,11], the rich collection of *C. difficile* methylomes and their genomes provides an unprecedented opportunity to jointly analyze its diversity of defense systems in a data-driven manner. In addition to R-M systems, we searched for evidence of additional systems: CRISPR-Cas, toxin-antitoxin (T-A), abortive infection (Abi) systems, bacteriophage exclusion (BREX)[12], prokaryotic Argonautes (pAgos)[13], DISARM[14], and a set of 10 recently-discovered defense systems[15] (Materials and Methods). Only T-A and CRISPR-Cas systems were ubiquitous in our genomes (Extended Data Fig. 1a, Supplementary Tables 3a-d), and all CRISPR-Cas systems detected were of Type-IB[16], consistent with earlier studies[17-19] (Extended Data Fig. 1b).

Different types of defense systems are expected to confer different degrees of protection against invading DNA. Also, under certain conditions, some defense systems may even facilitate genetic exchange between cells, as recently shown for CRISPR-Cas[20] and R-M systems[21]. Thus, we enquired how gene flux was distributed within our dataset (Materials and Methods, Extended Data Fig. 1a, Supplementary Tables 3e, f), and how is it associated with the multiple defense systems present in it. Specifically, to test the effect of

R-Ms, CRISPR-Cas, and T-As (the most abundant defense systems found across *C. difficile* strains) on gene flux, we built stepwise linear models to assess the role of each of these variables in explaining the variance of HGT and HR. Interestingly, we found a strong positive association between CRISPR spacer count and HGT/HR (Supplementary Table 3g, Extended Data Fig. 2). The increase in gene flux with spacer content, although somehow unexpected, could be reconciled if generalized transduction occurs in *C. difficile*[22,23], as recently shown for *Pectobacterium*[20]. R-M and T-A abundance had a less important explanatory role in the prevention of gene flux (Supplementary Table 3g).

Next, we made use of i) the unique information on R-M recognition motifs provided by SMRT-seq, and ii) the exceptional diversity of Type I R-M systems (as well as the near depletion of other types of complete systems) observed in our *C. difficile* dataset, to better understand their impact on phage target side avoidance and gene flux.

Restriction site avoidance is the most effective way to escape the action of R-M systems, and it has been predominantly studied for those belonging to Type II systems[24-27]. To investigate this, we selected representative members of the *Siphoviridae* and *Myoviridae* families and used Markov chain models to compute the number of observed and expected Type I R-M target sites accounting for oligonucleotide composition of each phage genome (Materials and Methods). We then used as a measure of genetic transfer, the number of recent HGT gains from the pattern of presence/absence of gene families in the species tree[28]. Our data suggest that *C. difficile* phages have generally evolved to reduce the number of several Type I recognition sites (Extended Data Fig. 3a). Concomitantly, we found an inverse linear trend between HGT and O/E ratios of Type I motif targets in phages (Spearman's ρ = -0.826, $P < 0.05$) (Extended Data Fig. 3b). This suggests a link between the frequencies at which different *C. difficile* genomes (or STs) are targeted by

phages, and the latter's capacity to underrepresent certain motifs targeted by the cell's R-M machinery.

Estimation of gene flux in *C. difficile* genomes

We used two measures of genetic transfer: the first based on the number of recent HGT gains and the second based on the number of HR events in the core-genome using two different programs (ClonalFrameML and Geneconv) (Materials and Methods). We identified 4,928 events of gene transfer in the *C. difficile* pan-genome (Supplementary Table 3e, Extended Data Fig. 1a). These events were very unevenly distributed across branches (and STs), from no events in CD_020265, to 530 in CD_020486. For HR, we found 403 and 270 events in the core-genome as given by Geneconv and CFML respectively (Supplementary Table 3e, Extended Data Fig. 1a). Even if the two programs provided different numbers of events, their results per genome were significantly correlated (Spearman's correlation = 0.60, $P < 10^{-3}$). To further complement the information on HR and HGT, we performed a one-way hierarchical clustering with the number of matches between spacers and currently known *Clostridium* phage sequences (Materials and Methods). We found very heterogeneous targeting profiles across isolates (Extended Data Fig. 1a, Supplementary Table 3f), which still correlated with HGT and HR (both with Spearman's correlation = 0.33, $P < 0.05$).

Core and pan-genome analysis of *C. difficile*

The core-genome contains 2,118 orthologous gene families (Extended Data Fig. 6b), corresponding to 64.6% of the smallest genome (*C. difficile* M120). Gene rarefaction analyses showed that the core-genome varies little with the addition of the last genomes (Extended Data Fig. 6b), suggesting that this estimate is robust. *C. difficile* has a large

pan-genome with a total of 8,246 gene families (Extended Data Fig. 6b). The spectrum of gene frequencies for the *C. difficile* pan-genome showed that the vast majority of gene families were either encoded in a few genomes (47.9% in four or less) or in most of them (37% in more than 41 genomes) (Extended Data Fig. 6c). Hence, *C. difficile* is characterized by a large and diverse pan-genome and low levels of genome conservation, which fall in the interval of previous estimates[29-31].

Inference of homologous recombination events by ClonalFrameML

HR is an evolutionary mechanism that takes place between highly similar sequences[32], and has been heralded as one of the evolutionary forces underlying the emergence of *C. difficile*'s pathogenicity[29,33]. This prompted us to examine whether HR could contribute to across-genome variation of CAAAAA motif sites located in the core-genome. We started by quantifying the diversity of gene repertoires (core- and pan-genome) in *C. difficile* (Extended Data Figs. 6b,c), and then inferred 277 recombination events at the phylogenetic tree tips (Supplementary Table 3e). To test whether HR contributes to across-genome variation of CAAAAA motif sites located in the core-genome, we also performed a systematic analysis of such events (Supplementary Figs. 6b-d, Supplementary Table 6g) and found that HR tracts indeed over-represent (O/E=1.40, $P < 10^{-3}$; Chi-square test) orthologous variable CAAAAA motif positions, while the core-genome without HR tracts underrepresents them (O/E=0.89, $P < 10^{-3}$; Chi-square test) (Extended Data Figs. 6e, f). While the detailed breakdown of HR events suggests great variation across genomes (Extended Data Fig. 6d), two primary HR peaks were prominent. The first (0.3-0.7 Mb in the core-genome) encompasses genes involved in motility and chemotaxis, such as those belonging to the *flg*, *fli*, and *flh* operons, as well as genes pertaining to ABC and PTS transport systems. The second (2.6-2.7 Mb in the core

genome) harbors genes pertaining to the S-layer (*slpA*, *cwp*, etc). Interestingly, both peaks match regions enriched for CAAAAA sites (Fig. 3a), suggesting a possible co-localization. The analysis showed that the recombination to mutation ratio (R/θ) is 0.162, the average length of recombined fragments (δ), is 159.4 bp, and the average distance between donor and recipient is 0.047. Thus, mutations are roughly 6.16 times more frequent than recombination, while the impact of recombination over mutation is 1.22 higher towards the evolution of these strains.

Supplementary references

1. Deakin LJ*, et al.* The *Clostridium difficile spo0A* gene is a persistence and transmission factor. *Infect. Immun.* **80**, 2704-2711 (2012).

2. Paredes-Sabja D, Shen A, Sorg JA. *Clostridium difficile* spore biology: sporulation, germination, and spore structural proteins. *Trends Microbiol.* **22**, 406-416 (2014).

3. Seekatz AM, Young VB. *Clostridium difficile* and the microbiota. *J. Clin. Invest.* **124**, 4182-4189 (2014).

4. Zidaric V, Rupnik M. Sporulation properties and antimicrobial susceptibility in endemic and rare *Clostridium difficile* PCR ribotypes. *Anaerobe* **39**, 183-188 (2016).

5. Collins J*, et al.* Dietary trehalose enhances virulence of epidemic *Clostridium difficile*. *Nature* **553**, 291-294 (2018).

6. Lanis JM, Barua S, Ballard JD. Variations in TcdB activity and the hypervirulence of emerging strains of *Clostridium difficile*. *PLoS Pathog.* **6**, e1001061 (2010).

7. Valiente E, Cairns MD, Wren BW. The *Clostridium difficile* PCR ribotype 027 lineage: a pathogen on the move. *Clin. Microbiol. Infect.* **20**, 396-404 (2014).

8. Lessa FC*, et al.* Burden of *Clostridium difficile* infection in the United States. *N. Engl. J. Med.* **372**, 825-834 (2015).

9. Clark TA*, et al.* Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biol.* **11**, 4 (2013).

10. Yaung SJ, Esvelt KM, Church GM. CRISPR/Cas9-mediated phage resistance is not impeded by the DNA modifications of phage T4. *PLoS ONE* **9**, e98811 (2014).

11. Dupuis ME, Villion M, Magadan AH, Moineau S. CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. *Nat. Commun.* **4**, 2087 (2013).

12. Goldfarb T*, et al.* BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.* **34**, 169-183 (2015).

13. Swarts DC*, et al.* DNA-guided DNA interference by a prokaryotic Argonaute. *Nature* **507**, 258-261 (2014).

14. Ofir G*, et al.* DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat. Microbiol.* **3**, 90-98 (2018).

15. Doron S*, et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, (2018).

16. Makarova KS*, et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**, 722-736 (2015).

17. Boudry P*, et al.* Function of the CRISPR-Cas system of the human pathogen *Clostridium difficile*. *MBio* **6**, e01112-01115 (2015).

18. Hargreaves KR, Flores CO, Lawley TD, Clokie MR. Abundant and diverse clustered regularly interspaced short palindromic repeat spacers in *Clostridium difficile* strains and prophages target multiple phage types within this pathogen. *MBio* **5**, e01045-01013 (2014).

19. Andersen JM, Shoup M, Robinson C, Britton R, Olsen KE, Barrangou R. CRISPR diversity and microevolution in *Clostridium difficile*. *Genome Biol. Evol.* **8**, 2841-2855 (2016).

20. Watson BNJ, Staals RHJ, Fineran PC. CRISPR-Cas-mediated phage resistance enhances horizontal gene transfer by transduction. *MBio* **9**, (2018).

21. Oliveira PH, Touchon M, Rocha EP. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl. Acad. Sci. USA* **113**, 5658-5663 (2016).

22. Hargreaves KR, Clokie MR. *Clostridium difficile* phages: still difficult? *Front. Microbiol.* **5**, 184 (2014).

23. Goh S, Hussain H, Chang BJ, Emmett W, Riley TV, Mullany P. Phage varphiC2 mediates transduction of Tn6215, encoding erythromycin resistance, between *Clostridium difficile* strains. *MBio* **4**, e00840-00813 (2013).

24. Rocha EP, Danchin A, Viari A. Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. *Genome Res.* **11**, 946-958 (2001).

25. Karlin S, Burge C, Campbell AM. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res.* **20**, 1363-1370 (1992).

26. Roberts GA*, et al.* Impact of target site distribution for Type I restriction enzymes on the evolution of methicillin-resistant *Staphylococcus aureus* (MRSA) populations. *Nucleic Acids Res.* **41**, 7472-7484 (2013).

27. Sharp PM. Molecular evolution of bacteriophages: evidence of selection against the recognition sites of host restriction enzymes. *Mol. Biol. Evol.* **3**, 75-83 (1986).

28. Csuros M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910-1912 (2010).

29. He M*, et al.* Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc. Natl. Acad. Sci. USA* **107**, 7527-7532 (2010).

30. Scaria J, Ponnala L, Janvilisri T, Yan W, Mueller LA, Chang YF. Analysis of ultra low genome conservation in *Clostridium difficile*. *PLoS ONE* **5**, e15147 (2010).

31. Forgetta V*, et al.* Fourteen-genome comparison identifies DNA markers for severe-disease-associated strains of *Clostridium difficile*. *J. Clin. Microbiol.* **49**, 2230-2238 (2011).

32. Vulic M, Dionisio F, Taddei F, Radman M. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc. Natl. Acad. Sci. USA* **94**, 9763-9767 (1997).

33. Yahara K*, et al.* The landscape of realized homologous recombination in pathogenic bacteria. *Mol. Biol. Evol.* **33**, 456-471 (2016).