nature microbiology

ARTICLES

https://doi.org/10.1038/s41564-019-0613-4

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[1,8], Robert Sebra[1], Harm van Bakel[1], Andrew Kasarskis[1], Rita Tamayo[3], Aimee Shen[2*] and Gang Fang[1*]

***Clostridioides* (formerly *Clostridium*) *difficile* is a leading cause of healthcare-associated infections. Although considerable progress has been made in the understanding of its genome, the epigenome of *C. difficile* and its functional impact has not been systematically explored. Here, we perform a comprehensive DNA methylome analysis of *C. difficile* using 36 human isolates and observe a high level of epigenomic diversity. We discovered an orphan DNA methyltransferase with a well-defined specificity, the corresponding gene of which is highly conserved across our dataset and in all of the approximately 300 global *C. difficile* genomes examined. Inactivation of the methyltransferase gene negatively impacts sporulation, a key step in *C. difficile* disease transmission, and these results are consistently supported by multiomics data, genetic experiments and a mouse colonization model. Further experimental and transcriptomic analyses suggest that epigenetic regulation is associated with cell length, biofilm formation and host colonization. These findings provide a unique epigenetic dimension to characterize medically relevant biological processes in this important pathogen. This study also provides a set of methods for comparative epigenomics and integrative analysis, which we expect to be broadly applicable to bacterial epigenomic studies.**

Clostridioides difficile is a spore-forming Gram-positive obligate anaerobe and the leading cause of nosocomial antibiotic-associated disease in the developed world[1] (Supplementary Notes). Despite the substantial progress that has been achieved in the understanding of *C. difficile* physiology, genetics and genomic evolution[2,3], the roles played by epigenetic factors—namely DNA methylation—have not been systematically studied[4–6]. In the bacterial kingdom, there are three major forms of DNA methylation: N^6-methyladenine (6mA; the most prevalent form representing ~80%), N^4-methylcytosine (4mC) and 5-methylcytosine (5mC). Increasing evidence suggests that DNA methylation regulates a number of biological processes, including DNA replication and repair, cell cycle, chromosome segregation and gene expression[7–13]. Efficient high-resolution mapping of bacterial DNA-methylation events has only recently become possible with the advent of single-molecule real-time sequencing (SMRT-seq)[14,15]. This technique enabled the characterization of the first bacterial methylomes[16,17] and, since then, more than 2,200 (as of September 2019) have been mapped, heralding a new era of bacterial epigenomics[18].

Here we mapped and characterized the DNA methylomes of 36 human *C. difficile* isolates using SMRT-seq and comparative epigenomics. We observed substantial epigenomic diversity across *C. difficile* isolates, as well as the presence of a highly conserved methyltransferase (MTase). Inactivation of this MTase had a functional impact on sporulation, a key step in the transmission of *C. difficile*. Further experimental and integrative transcriptomic analysis suggested that epigenetic regulation by DNA methylation also modulates the cell length, host colonization and biofilm formation of *C. difficile*. These discoveries are expected to stimulate future investigations along a new epigenetic dimension to characterize and potentially repress medically relevant biological processes in this important pathogen.

Results

Methylome analysis reveals great epigenomic diversity in *C. difficile*. From an ongoing Pathogen Surveillance Program at Mount Sinai Medical Center, 36 *C. difficile* isolates were collected from faecal samples of infected patients (Supplementary Table 1). A total of 15 different multilocus sequence types (STs) belonging to clades 1 (human and animal, HA1) and 2 (hypervirulent or epidemic)[19] are represented in our dataset (Fig. 1a). Using SMRT-seq with long-library-size selection, de novo genome assembly was achieved at

[1]Department of Genetics and Genomic Sciences, Institute for Genomics and Multiscale Biology, Mount Sinai School of Medicine, New York, NY, USA. [2]Department of Molecular Biology and Microbiology, Tufts University School of Medicine, Boston, MA, USA. [3]Department of Microbiology and Immunology, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, NC, USA. [4]Microbial Evolutionary Genomics, Institut Pasteur, Paris, France. [5]CNRS, UMR3525, Paris, France. [6]Department of Medicine, Division of Infectious Diseases, Mount Sinai School of Medicine, New York, NY, USA. [7]Department of Pharmacological Sciences and Department of Oncological Sciences, Mount Sinai School of Medicine, New York, NY, USA. [8]Sema4, Stamford, CT, USA. *e-mail: aimee.shen@tufts.edu; gang.fang@mssm.edu

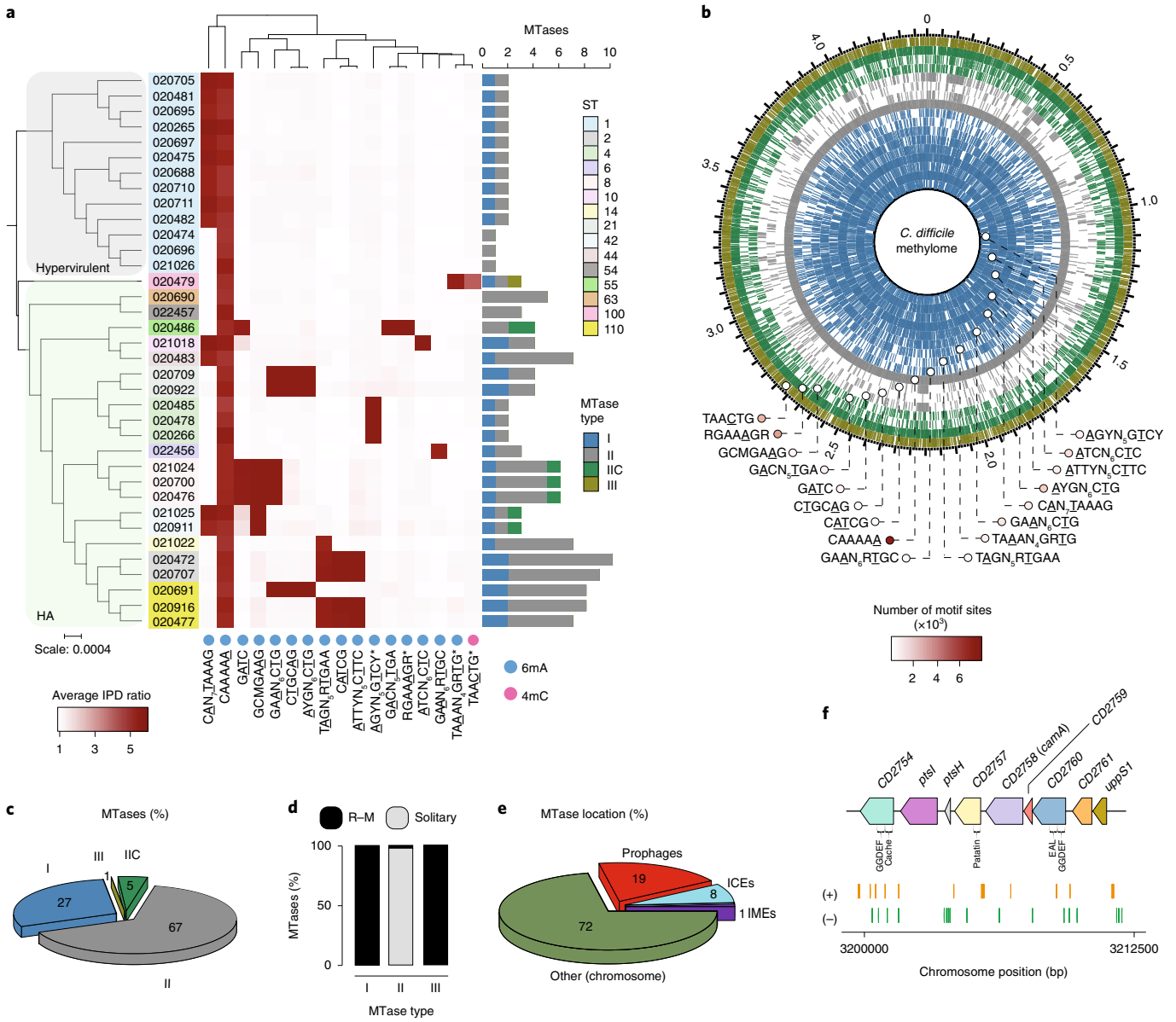


Fig. 1 | The methylomes of the 36 strains of *C. difficile*. **a**, A phylogenetic tree of the 36 *C. difficile* strains coloured by clade (hypervirulent, human and animal (HA) associated) and ST. A heat map of the landscape of methylated motifs per genome, and their average interpulse duration (IPD) ratio is also shown. The asterisks indicate new motifs that were not previously listed in the reference database REBASE. Methylated bases are underlined. The CAAAAA motif was consistently methylated across isolates. The bar plot indicates the number and types of active MTases detected per genome. In type IIC systems, MTase and restriction endonuclease (REase) are encoded in the same polypeptide. **b**, The *C. difficile* methylome. The positions of all of the methylation motif sites in the reference genome of *C. difficile* 630 are indicated, coloured according to MTase type. The average motif occurrences per genome (across the 36 isolates) are also indicated. **c**, The percentage of MTases detected according to type. **d**, The percentage of MTases pertaining to complete R–M systems or without cognate REase (solitary). **e**, Breakdown of MTases by location—integrative mobile elements (IMEs), integrative conjugative elements (ICEs), prophages and other (within the chromosome). No hits were obtained in plasmids. **f**, Immediate genomic context of *camA*. The example shown (including coordinates) refers to the reference genome of *C. difficile* 630. The plus and minus signs indicate the sense and antisense strands, respectively. The vertical bars indicate the distribution of the CAAAAA motif. CD2754, a phosphodiesterase with a GGDEF domain (PFAM PF00990) and a cache domain (PF02743); *ptsI* and *ptsH* belong to a phosphotransferase (PTS) system; CD2757, patatin-like phospholipase (PF01734); CD2758 (*camA*), type II MTase; CD2759, Rrf2-type transcriptional regulator; CD2760, phosphodiesterase with a GGDEF domain and a conserved EAL domain (PF00563); CD2761, *N*-acetylmuramoyl-L-alanine amidase; *uppS1*, undecaprenyl diphosphate synthase. The genomic context of *camA* is largely conserved across strains, located approximately 25 kb upstream of the S-layer biogenesis locus (Extended Data Fig. 4c,d). Several of the genes that flank *camA* (in addition to *camA* itself) are part of the *C. difficile* core genome (see below), suggesting that they may have biological functions that are fundamental to *C. difficile*.

a high quality (Supplementary Table 1). Methylation motifs were found using the SMRTportal protocol. We found a total of 17 unique high-quality methylation motifs in the 36 genomes (an average of 2.6 motifs per genome; Fig. 1a, Supplementary Table 2a). The large majority of target motifs were of the 6mA type, one motif (TAACTG; methylated bases are underlined) belonged to the 4mC type and no 5mC motifs were detected with confidence (Supplementary Notes). As with most bacterial methylomes, more than 95% of the 6mA and 4mC motif sites were methylated (Fig. 1b, Supplementary Table 2a).

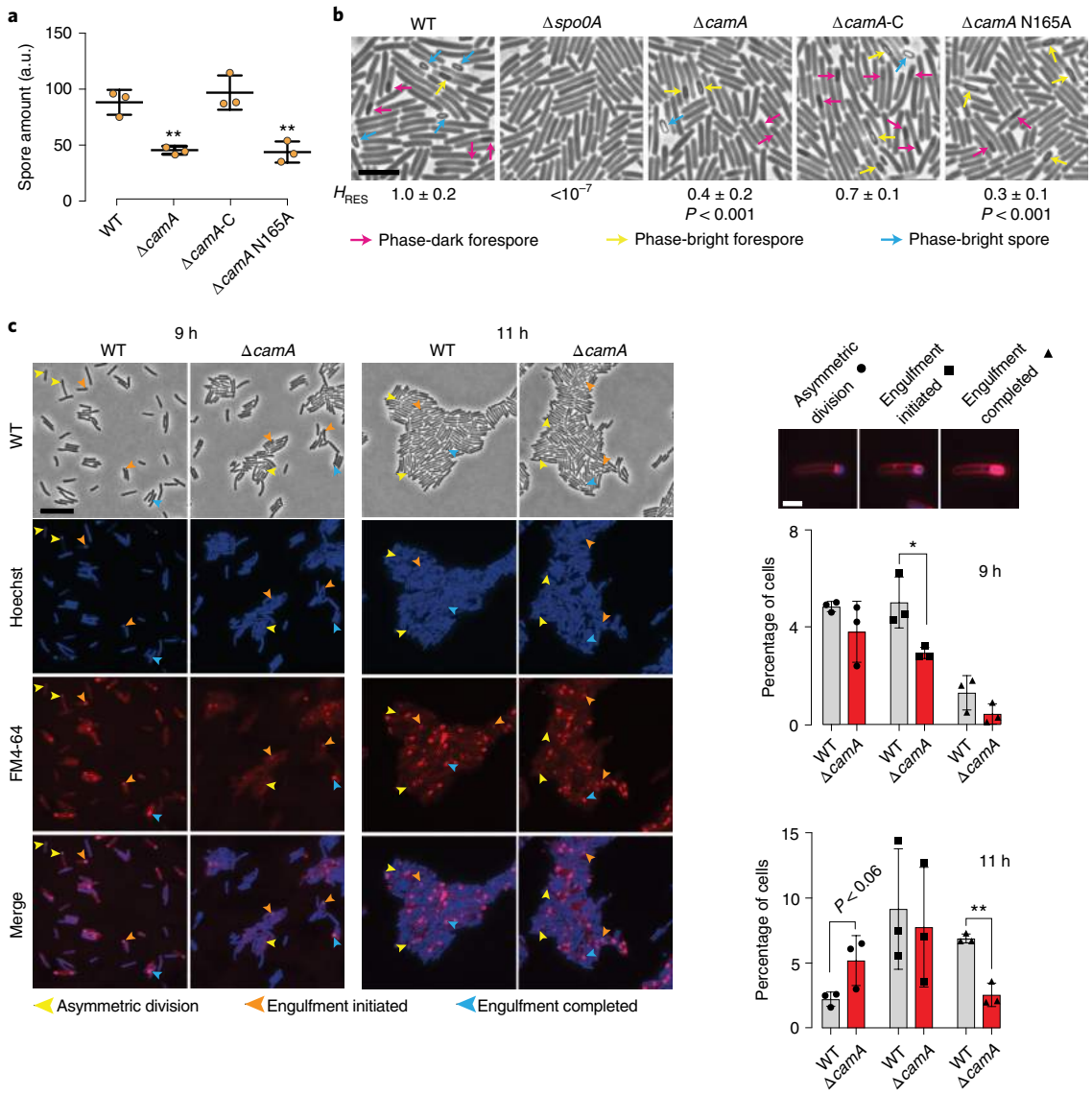


Fig. 2 | CamA modulates sporulation levels in *C. difficile*. a, Spore purification efficiencies obtained from sporulating cells; $n = 3$ independent spore preparations. $^{**}P < 10^{-2}$; the statistical analysis was performed using one-way ANOVA and Tukey's test. The spore yield was determined by measuring the optical density at 600 nm (OD_{600}) of the resulting spore preparations and correcting for the volume of resuspension water. Data are mean ± s.d. **b**, Phase-contrast microscopy after 20 h of sporulation induction. The Δ*spo0A* strain was used as a negative control because it does not initiate sporulation[44]. Immature phase-dark forespores are indicated by pink arrows, and mature phase-bright forespores and free spores are indicated by yellow and blue arrows, respectively. Scale bar, 5 µm. Heat-resistance (H_{RES}) efficiency data are mean ± s.d. of $n = 3$ independent replicates. **c**, Morphological analysis of WT and Δ*camA* cells using fluorescent stains comparing 9 h and 11 h after sporulation induction. The polar septum formed during asymmetric division is visible using FM4-64 membrane staining, whereas the chromosome that is pumped into the forespore after polar-septum formation can be seen as a bright focus using Hoechst DNA staining. FM4-64 staining enables the visualization of engulfing membranes. As the mother-cell-derived membrane fully encircles the forespore-derived membrane, the FM4-64 signal becomes more intense around the forespore. When these membranes undergo fission, the forespore becomes fully suspended in the cytosol of the mother cell, and both stains are excluded. The yellow arrows indicate cells that are undergoing asymmetric division (indicated by a flat polar septum); the orange arrows indicate cells that are in the process of engulfment (indicated by a curved polar septum); and blue arrows indicate cells that have completed engulfment (indicated by bright membrane staining fully surrounding the forespore). Scale bar, 10 µm. The bar plots indicate the percentage of sporulating cells at different stages of spore assembly in both WT and Δ*camA* cells. Data are mean ± s.d. of $n = 3$ independent replicates. Images above bar plots show examples of each spore assembly stage; scale bar, 2 µm. A total of 3,747 (WT, 9 h), 3,879 (Δ*camA*, 9 h), 4,960 (WT, 11 h) and 4,650 (Δ*camA*, 11 h) cells were screened. $^{*}P \leq 0.05$; the statistical analysis was performed using two-tailed unpaired Student's *t*-tests.

Genomes pertaining to the same ST tend to have more similar sets of methylation motifs relative to those from different STs. Those genomes belonging to ST-2, ST-8, ST-21 and ST-110 showed the highest motif diversities. One 6mA motif, CAAAAA, was present across all of the genomes; we therefore hypothesized that 6mA methylation events at this motif and its corresponding MTase have an important and conserved function in *C. difficile*.

A DNA methyltransferase and its target motif are ubiquitous in *C. difficile*. Motivated by the consistent presence of the methylation

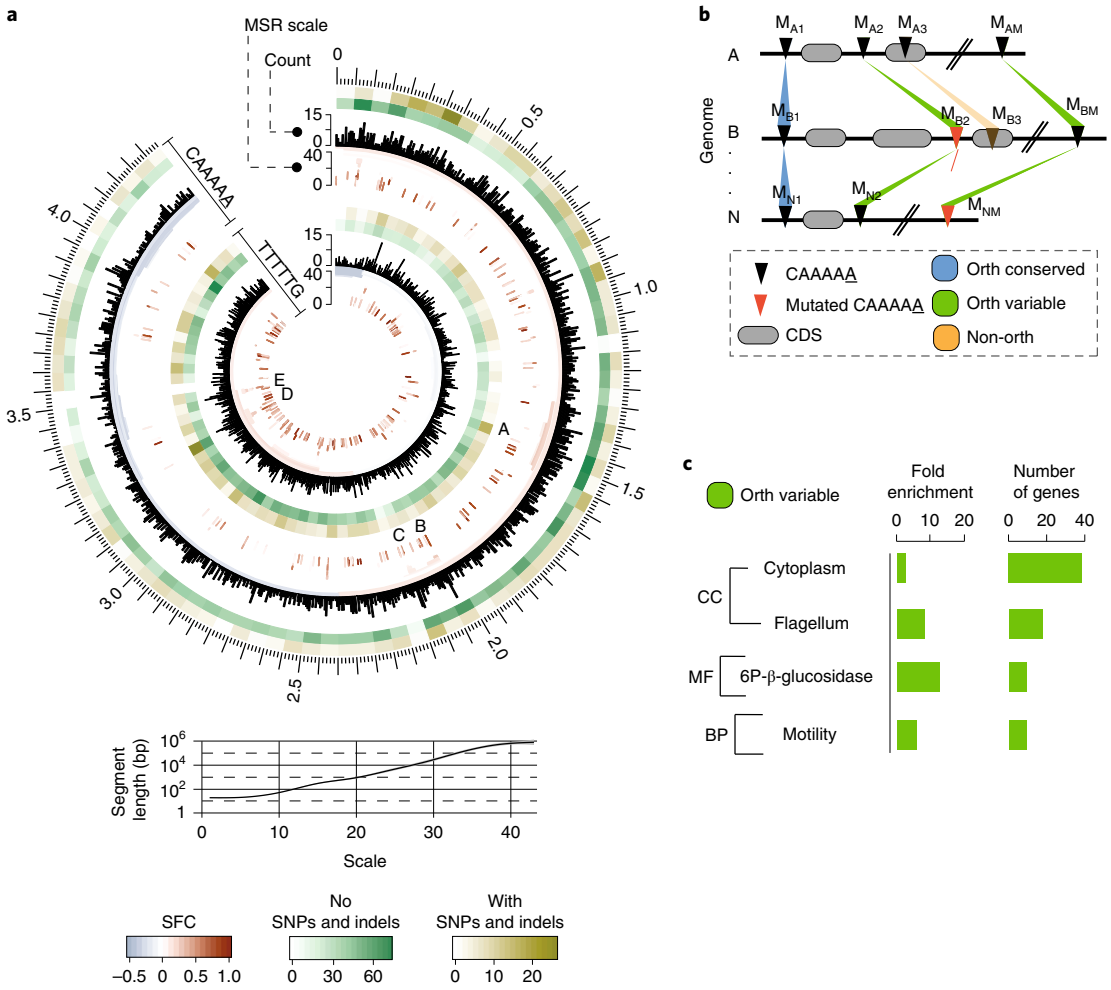


Fig. 3 | Abundance, distribution and conservation of CAAAAA motif sites. a, The distribution of CAAAAA sites in both strands of the reference *C. difficile* 630 genome and corresponding genomic signal obtained by MSR. In brief, MSR uses wavelet transformation to examine the chromosome at a succession of increasing length scales by testing for enrichment or depletion of a given genomic signal. Whereas scale values of less than 10 are typically associated with regions of less than 100 bp, genomic regions enriched for CAAAAA sites at scale values of more than 20 correspond to segments larger than 1 kb (that is, gene and operon scale). The letters (A–E) represent regions with a particularly high abundance of CAAAAA motif sites, including genes related to sporulation (such as *spo0A*, *spoIIAA-AH*, *spoIVB* and *sigK*), membrane transport (PTS and ABC-type systems), transcriptional regulation (such as *iscR* and *fur*) and coding for multiple cell wall proteins (Supplementary Table 6d). The relationship between MSR scale and segment length is also shown. The significant fold change (SFC) corresponds to the fold change ($\log_2$-transformed ratio) between observed and randomly expected overlap, statistically significant at $P = 10^{-6}$ on the basis of the *Z*-test. Heat-map layers correspond to the number of orthologous conserved (no SNPs or indels; green) and orthologous variable (with SNPs and/or indels) CAAAAA motif positions. **b**, Whole-genome alignment of 37 *C. difficile* genomes (36 isolates and *C. difficile* 630 as reference) was performed using Mauve. We defined an orthologous (Orth) occurrence of the CAAAAA motif (black triangles) if an exact match to the motif was present in each of the 37 genomes (conserved; blue) or if at least one motif (and a maximum of $n-1$, being n the number of genomes) contained positional polymorphisms (maximum of two SNPs or indels per motif; variable; green). Non-orthologous CAAAAA positions are indicated as orange regions. The results are shown in **a** as heat maps. The numbering scheme is based on mapping location. **c**, DAVID enrichment analysis of genes containing intragenic and regulatory (100 bp upstream the start codon) orthologous variable CAAAAA motif sites. Genes that were found to over-represent orthologous variable CAAAAA positions include cytoplasm-related genes (such as *pheA*, *fdhD*, *ogt1* and *spoIVA*) and motility-related genes (such as *fliZ*, *fliN*, *fliM* and *flgL*). CC, cellular component; MF, molecular function; BP, biological process. Single categories were considered to be significantly enriched at $P < 0.05$ and correspond to 73 out of a total of 617 genes analysed; FDR-corrected *P* values were calculated using one-tailed Fisher's exact tests.

motif CAAAAA across all of the *C. difficile* isolates, we proceeded to examine the encoded MTases. We identified a total of 139 MTase genes (an average of 3.9 per genome; Fig. 1a, Supplementary Table 2b) that represent all of the four major types[20] and appear either in a solitary context or within restriction–modification (R–M) systems (Fig. 1c–e, Supplementary Table 2b–d). We also found multiple additional defence systems (such as abortive infection systems, CRISPR–Cas and toxin–antitoxin) and performed an integrative analysis with R–M systems in relation to host defence and gene flux (Extended Data Figs. 1 and 2, Supplementary Table 3a–g), such as that resulting from phage infection (Extended Data Fig. 3, Supplementary Notes).

Consistent with the presence of a highly conserved CAAAAA motif, we identified a type II 6mA solitary DNA MTase (577 amino acids) that is present across isolates (Fig. 1f, Supplementary Table 2b, Supplementary Notes) and is responsible for methylation of the CAAAAA motif. This MTase is encoded by *CD2758* in the reference strain *C. difficile* 630 (refs. [2,21]). Here we named CD2758 CamA

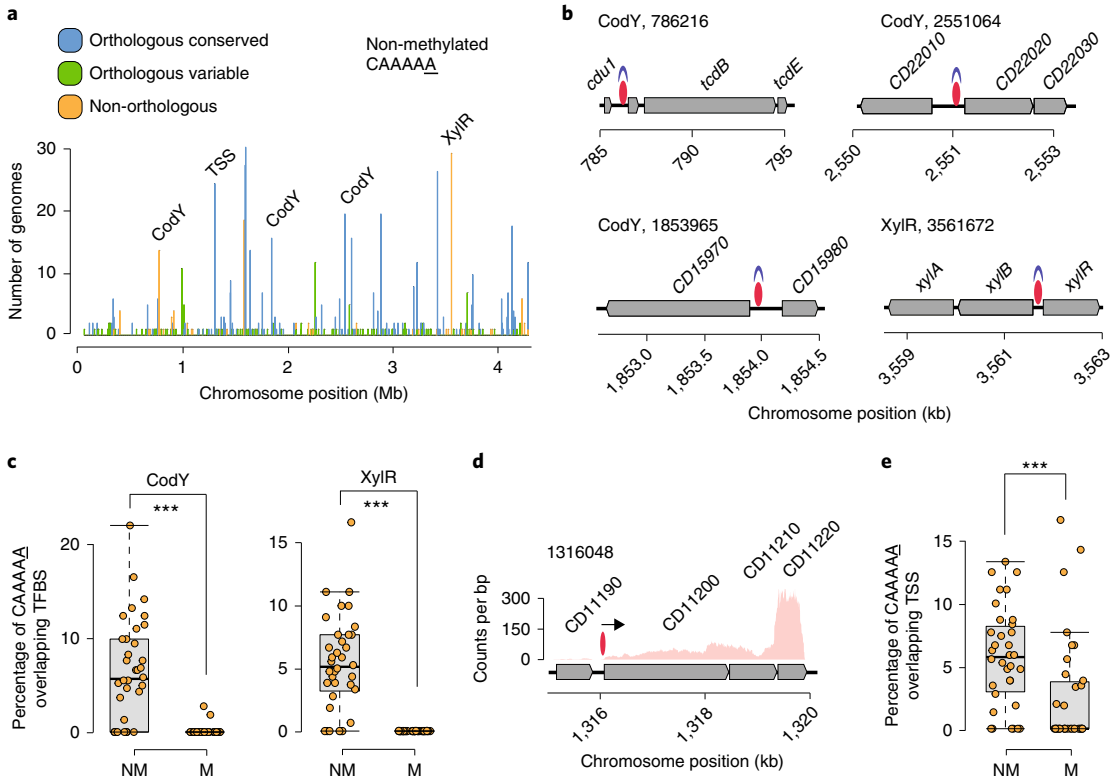


Fig. 4 | The distribution of non-methylated CAAAAA motif sites and their overlap with TFBSs and TSSs. a, The number of *C. difficile* isolates in which non-methylated CAAAAA motif sites were detected at a given chromosome position (coordinates are relative to the reference genome of *C. difficile* 630). Peak colours correspond to orthologous (conserved and variant) and non-orthologous CAAAAA positions. We found that some of the major peaks of non-methylated CAAAAA positions overlapped with TFBSs (such as CodY and XylR) and TSSs. **b**, Genetic regions for which overlap was observed between highly conserved non-methylated CAAAAA motif sites (red ovals) and TFs (CodY and XylR, shown in blue). Other examples of conserved non-methylated CAAAAA motif sites are provided in Extended Data Fig. 7b. **c**, The percentage of CAAAAA motif sites (non-methylated (NM) and methylated (M)) that overlap CodY and XylR binding sites for each of the $n=36$ *C. difficile* isolates. $^{***}P<10^{-3}$. **d**, An example of a chromosomal region in which non-methylated CAAAAA motifs overlap a TSS (arrow). **e**, The percentage of CAAAAA motifs (non-methylated and methylated) that overlap TSSs for each of the $n=36$ *C. difficile* isolates. For the box plots, the centre line indicates the median value, the boxes indicate the 25th and 75th quartiles, and the whiskers indicate 1.5× the interquartile range. For **c** and **e**, the statistical analysis was performed using one-sided Mann–Whitney–Wilcoxon rank-sum tests with continuity correction.

(*C. difficile* adenine methyltransferase A). Its ubiquity was not restricted to the 36 isolates, as we were able to retrieve orthologues in a list of around 300 global *C. difficile* isolates from GenBank (Supplementary Table 4). REBASE also showed functional orthologues of *camA* in only a very small number of other *Clostridiales* and *Fusobacteriales* (Extended Data Fig. 4), suggesting that this MTase is fairly unique to *C. difficile*.

Inactivation of *camA* reduces sporulation levels in vitro. Given the critical role of sporulation in the persistence and dissemination of *C. difficile* in humans and hospital settings[22], we decided to test whether *camA* inactivation could reduce spore purification efficiencies in the 630 strain as previously suggested for its homologue in the 027 isolate R20291 (ref. [23]). We constructed an in-frame deletion in this gene (Δ*camA*) and complemented it with either wild-type (WT) *camA* (Δ*camA*–C) or a variant encoding a catalytic site mutation (N165A) of the MTase (Δ*camA* N165A) (Extended Data Fig. 5a, Supplementary Table 5a,b). We observed that spore purification efficiencies decreased by around 50% in the mutant relative to the WT (Fig. 2a). Complementation of Δ*camA* with the WT, but not the catalytic mutant, restored spore purification efficiencies to values that are similar to those observed in the WT cells (Fig. 2a, Supplementary Table 5c). No differences in growth were observed between the WT and mutant strains (Extended Data Fig. 5b). Thus, this complementation experiment suggests that the loss of methylation events by CamA, rather than the loss of non-catalytic roles of this protein, leads to the decrease in spore yield.

The diminished efficiencies in spore purification observed in the Δ*camA* mutants could be due to a reduction in the number of cells that induce sporulation or defects in spore assembly[24]. Visual inspection of samples before and after spore purification on a density gradient revealed qualitatively lower levels of mature phase-bright spores (Extended Data Fig. 5c). As purified WT and Δ*camA* spores had similar levels of chloroform resistance and germinated with similar efficiency (Extended Data Fig. 5d,e), the reduced spore purification efficiencies of the MTase mutants probably reflect a defect in sporulation initiation rather than the sporulation process itself. Accordingly, we observed that fewer Δ*camA* cells were sporulating in phase-contrast microscopy analyses relative to WT cells (Fig. 2b).

To gain insights into the sporulation stage that is affected by the loss of CamA, we quantified the number of sporulating cells at different stages of spore assembly (Fig. 2c). Although similar numbers of WT and Δ*camA* cells were observed at asymmetric division (the first morphological stage of sporulation) 9 h after sporulation induction, 50% fewer Δ*camA* cells had initiated engulfment. Furthermore, around twofold more Δ*camA* cells were at asymmetric division relative to WT cells 11 h after sporulation induction,

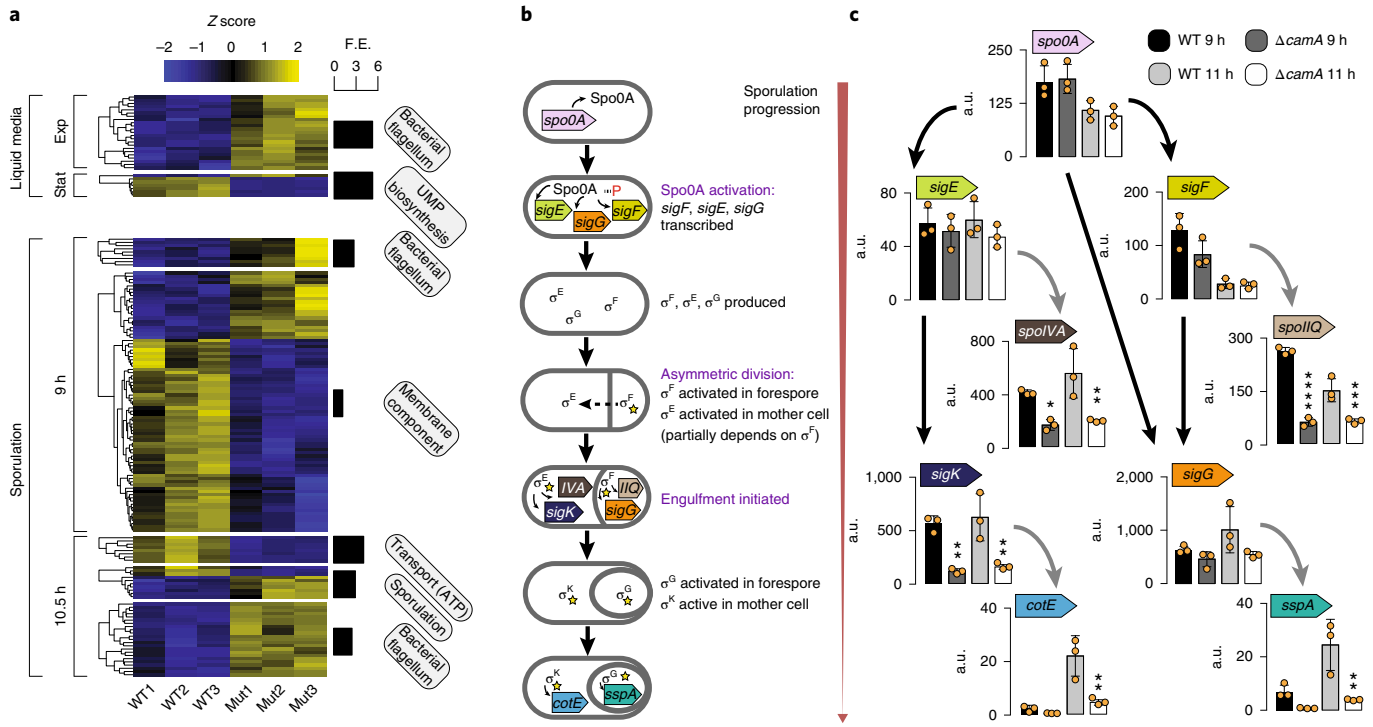


Fig. 5 | Gene-expression analysis. a, A heat map of 161 genes from three replicates of *C. difficile* 630 compared with an equal number of replicates of *C. difficile* 630 Δ*camA* that are enriched for the Gene Ontology (GO) terms shown in the boxes and described in Supplementary Table 8c. The *Z* score reflects the degree of downregulation (*Z* score < 0) or upregulation (*Z* score > 0), computed by subtracting the mean of the log-transformed expression values and dividing by the s.d. for each gene over all of the samples scored. FE, fold enrichment; Exp, exponential growth stage; Stat, stationary growth stage. **b**, Schematic of the sequence of sporulation sigma factor gene transcription and protein activation coupled to morphological changes during sporulation. Activated Spo0A induces the expression of genes encoding σ^F, σ^E and σ^G as well as factors required for asymmetric division and the post-translational activation of the early stage sporulation sigma factors, σ^F and σ^E. σ^F is the first sporulation-specific sigma factor to be fully activated and it becomes active in the forespore only after asymmetric division is completed[109]. Activated σ^F subsequently induces the transcription of genes of which the products mediate σ^G activation in the forespore and partially mediate σ^E activation in the mother cell[35]. Activated σ^E induces the transcription of *sigK*[33] and factors required for the excision of a prophage-like element from the *sigK* gene[110]. *C. difficile* sporulation is therefore controlled by a transcriptional hierarchy that is coupled to morphological events such that downstream sigma factors (σ^G and σ^K) depend on the activation of upstream sigma factors (σ^F and σ^E). **c**, A comparison of relative transcript levels in WT and Δ*camA* cells as determined by RT-qPCR for sporulation sigma factor genes and representative genes in the regulons of sporulation-specific sigma factors at 9 h and 11 h after sporulation induction (a separate set of $n = 3$ RNA sample replicates was used). Note that the primers for *sigK* amplify a region before the *sigK* excision site[110]. Data are mean ± s.d. $^*P < 0.05$, $^{**}P < 10^{-2}$, $^{***}P < 10^{-3}$, $^{****}P < 10^{-4}$; statistical significance was determined using one-way ANOVA and Tukey's test for multiple comparisons.

whereas 50% fewer Δ*camA* cells had completed engulfment compared with the WT cells. As similar numbers of sporulating cells were observed between WT and Δ*camA* at 11 h, the sporulation defect of Δ*camA* cells seems to arise because fewer cells progress beyond asymmetric division, rather than due to a defect in sporulation induction.

To confirm that the loss of CamA leads to a decrease in the number of cells that produce functional spores, we compared the ability of Δ*camA* to form heat-resistant spores that are capable of germinating and outgrowing using a heat-resistance assay[25]. The Δ*camA* mutant and the catalytic-mutant complementation strain produced approximately 50% fewer heat-resistant spores than the WT and WT-complementation strains (Extended Data Fig. 5e). Taken together, these findings suggest that CAAAAA̲ methylation enhances sporulation in vitro. This functional difference prompted us to perform a comprehensive methylome and transcriptome analysis of WT and Δ*camA* strains.

Comparative analysis of CAAAAA̲ sites across *C. difficile* genomes. The *C. difficile* genome has an average of 7,721 CAAAAA̲ motif sites (Supplementary Table 6a). Adjusting for the *k*-mer frequency of the AT-rich *C. difficile* genome (70.9%) using Markov models[26], CAAAAA̲ motif sites are significantly under-represented in intragenic regions (Extended Data Fig. 6a, Supplementary Table 6a,b). To evaluate whether specific chromosomal regions are enriched or depleted for this motif, we used a multiscale signal representation (MSR) approach[27]. We observed strong enrichment for CAAAAA̲ sites within genes related to sporulation, membrane transport, transcriptional regulation and coding for multiple cell wall proteins (Fig. 3a, Supplementary Table 6c,d).

To further characterize CAAAAA̲ motif sites, we categorized them on the basis of their positional conservation across genomes. We performed whole-genome alignment of the isolates and classified each motif position in the alignment as follows: (1) conserved orthologous (devoid of single-nucleotide polymorphisms (SNPs) or indels); (2) variable orthologous (in which at least one genome contains a SNP or indel); and (3) non-orthologous (Fig. 3b, Supplementary Data 1). We found a total of 5,828 conserved orthologous motif positions, 1,050 variable orthologous positions and an average of 843 non-orthologous positions per genome (Supplementary Table 6e). Among orthologous positions, the variable positions contribute to variations at CAAAAA̲ sites across genomes with subsequent methylation abrogation (Supplementary Table 6f). Such across-genome variation seems to be at least partially

fuelled by events of homologous recombination (Extended Data Fig. 6b–f, Supplementary Table 6g). Finally, DAVID gene enrichment analysis[28] found that cytoplasm- and motility-related genes over-represent orthologous variable CAAAAA positions (Fig. 3c). The very large number and dispersion of conserved orthologous positions precluded a similar functional analysis. Collectively, genome-wide distribution and across-genome comparative analyses suggest that CAAAAA sites are enriched in regions that harbour genes related to sporulation and colonization and that orthologous variable CAAAAA positions are enriched in regions that harbour cytoplasm- and motility-related genes.

Non-methylated CAAAAA motif sites are enriched in regulatory elements. The on/off switch of DNA methylation in a bacterial cell can contribute to epigenetic regulation as a result of competitive binding between DNA MTases and other DNA-binding proteins (such as transcription factors (TFs)) as previously described[12,29–31]. Previous bacterial methylome studies that analysed one or few genomes had insufficient statistical power to perform a systematic interrogation of non-methylated motifs sites[16]. Building on our collection of *C. difficile* methylomes, we performed a systematic detection and analysis of non-methylated CAAAAA sites and found an average of 21.5 of such sites per genome (Extended Data Fig. 7a, Supplementary Table 7a). We found that non-methylated motif sites were dispersed throughout the full length of the *C. difficile* chromosome, yet were over-represented in orthologous variable and non-orthologous CAAAAA positions (observed/expected, 1.51 and 1.49, respectively) and under-represented in orthologous conserved CAAAAA positions (observed/expected, 0.84; all $P<10^{-4}$; χ^2 test). This is consistent with the idea that variable positions are more likely to be non-methylated to provide breadth of expression variation. Most of the non-methylated positions (85.4% of 245) failed to conserve such status in more than three genomes at orthologous positions, whereas a small percentage of positions (5.5%) remained non-methylated in at least one-third of the isolates, suggesting that competitive protein binding is expected to be more active in certain genomic regions (Fig. 4a).

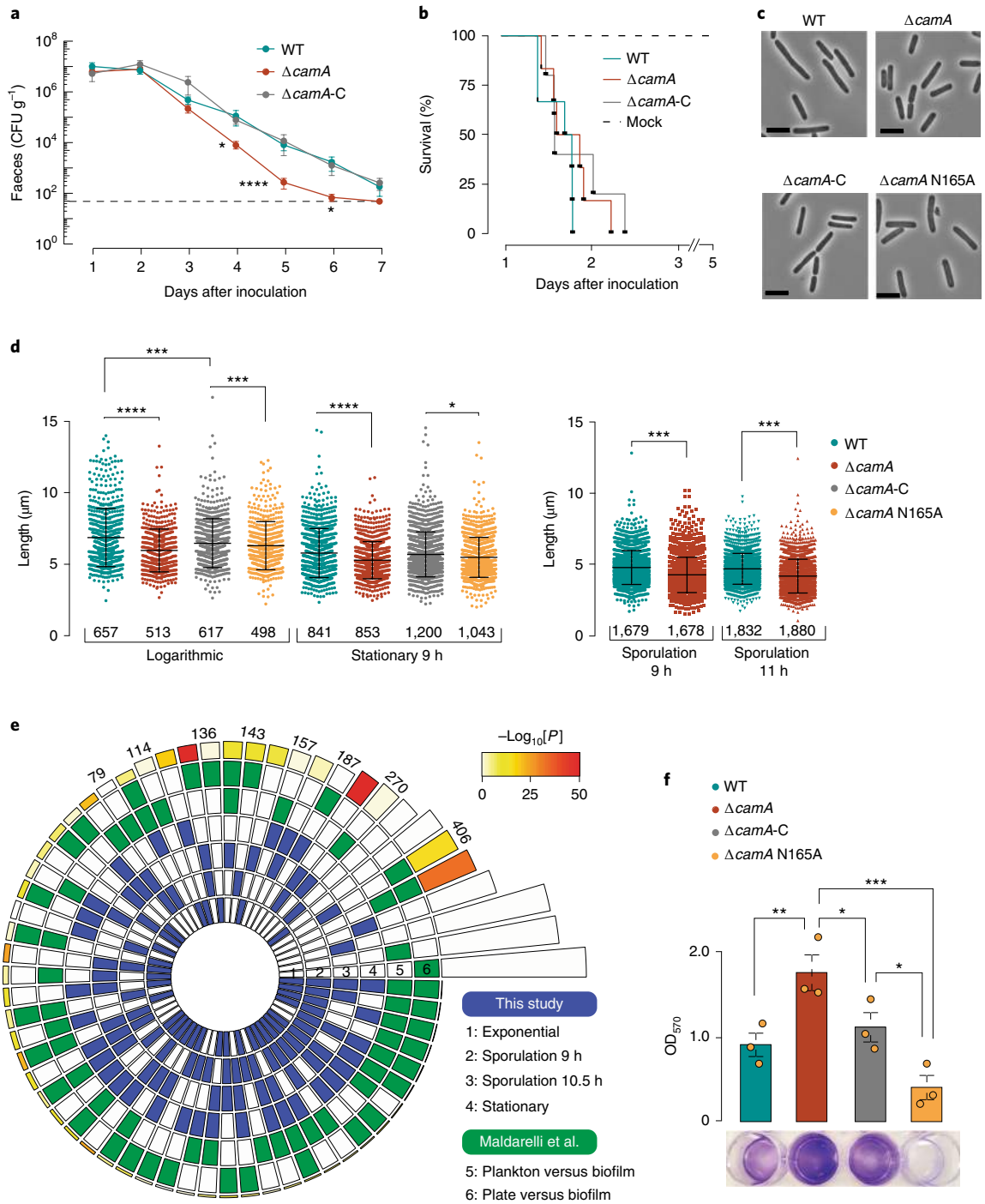
The non-methylated CAAAAA positions detected across the *C. difficile* genomes enabled us to perform a systematic search for evidence of overlap between the CAAAAA motifs and TF binding sites (TFBSs) and transcription start sites (TSSs). First, we found overlaps between prominent peaks of non-methylated CAAAAA positions and the TFBSs of CodY and XylR (Fig. 4a,b, Extended Data Fig. 7b, Supplementary Table 7b,c). Performing the analysis at the genome level, both CodY and XylR binding sites showed significant enrichment ($P<10^{-3}$, Mann–Whitney–Wilcoxon test) for non-methylated CAAAAA (Fig. 4c, Extended Data Fig. 7c). Second, using TSSs reconstructed from RNA-sequencing (RNA-seq) data coverage, we found a similar genome-level enrichment for non-methylated CAAAAA sites (Fig. 4d,e, Extended Data Fig. 7d,e, Supplementary Table 7d; $P<10^{-3}$, Mann–Whitney–Wilcoxon test). Thus, these results demonstrate the occurrence of an on/off epigenetic switch of CAAAAA sites that preferentially overlaps with putative TFBSs and TSSs.

Loss of CAAAAA methylation impacts the transcription of multiple gene categories, including sporulation. To study the functional importance of methylation at CAAAAA sites, we used RNA-seq to compare the transcriptome of WT *C. difficile* 630 with that of Δ*camA* both in liquid medium (exponential and stationary growth stage) and after sporulation induction (9 h and 10.5 h; Extended Data Fig. 8, Supplementary Table 8a, Supplementary Data 2). Of the 3,896 genes annotated in *C. difficile* 630, 36–361 (0.9–9.3%, depending on the time point) were differentially expressed (DE) at a 5% false discovery rate (FDR) and $|\log_2[FC]|>1$ (twofold change in gene expression; Fig. 5a, Supplementary Table 8b–d). DE genes in Δ*camA* cells relative to the WT showed significant enrichment in CAAAAA motif sites compared with non-DE genes ($P<10^{-2}$, Mann–Whitney–Wilcoxon test) in broth culture, and a qualitatively similar trend was also observed during sporulation (Extended Data Fig. 9a). Consistent with our finding that the loss of CamA reduces spore formation, the transcriptome analyses revealed that 118 and 120 genes that have previously identified to be induced during sporulation[32,33] were expressed at ≥50% reduced levels in Δ*camA* cells relative to WT cells at 9 h and 10.5 h, respectively (Supplementary Table 8b).

The transcriptional program that mediates sporulation in *C. difficile* is controlled by a master transcriptional activator, Spo0A, and four sporulation-specific sigma factors, σ^F, σ^E, σ^G and σ^K. These factors activate distinct regulons that ultimately lead to the assembly of functional spores[34,35] (Fig. 5b); the early acting sigma factors, σ^F and σ^E, are required for the activity of the later-acting sigma factors, σ^G and σ^K, respectively. A transcriptional hierarchy therefore governs sporulation in *C. difficile* with downstream factors depending on the activation of upstream sigma factors. As genes in the regulons of all four sporulation-specific sigma factors were underexpressed in Δ*camA* cells relative to the WT, whereas a relatively small subset of Spo0A regulon genes exhibited this pattern of regulation (Extended Data Fig. 9b, Supplementary Table 8e), loss of CamA probably affects early events during sporulation.

Fig. 6 | In vivo and additional functional impacts of the Δ*camA* mutation. a, Kinetics of infection in antibiotic-treated mice ($n=12$) after treatment with a sub-lethal inoculation (10^5 spores) of WT *C. difficile* 630Δ*erm*, MTase mutant Δ*camA* and complement Δ*camA*-C. When inoculated with spontaneous erythromycin sensitive derivative (630Δ*erm*) strains, antibiotic-treated mice do not typically develop fulminant disease and instead serve as a model of intestinal colonization and persistence by *C. difficile*[105,106,111]. The dotted line indicates the limit of detection. Data are mean ± s.e.m.; $\log_{10}$-transformed data from each time point were analysed by ANOVA for each time point. **b**, Kaplan–Meier survival curves for clindamycin-treated golden Syrian hamsters ($n=6$) infected with 10^3 spores of WT *C. difficile* 630Δ*erm*, Δ*camA* or complement Δ*camA*-C. **c**, Representative phase-contrast images ($n=3$ independent biological replicates) of vegetative WT, Δ*camA*, Δ*camA*-C and Δ*camA* N165A cells. Scale bars, 5 µm. **d**, Comparison of cell length. Data are mean ± s.d. of $n=3$ independent biological replicates (the exact numbers of cells measured are indicated in the figure). The statistical analysis was performed using one-way ANOVA and Tukey's test for multiple comparisons. **e**, Significance of overlap between multiple datasets of DE genes. Comparisons were performed between DE genes called in this study for each time point (blue, $n=1,537$) and those from Maldarelli et al.[40] (green, $n=1,735$). The latter corresponds to *C. difficile* DE genes in conditions that favour biofilm formation compared with growth on a plate or planktonic form. Colour intensities of the outermost layer represent the *P* value significance of the intersections (3,896 genes used as background). The height of the corresponding bars is proportional to the number of common genes in the intersection (shown for pairwise comparisons across different studies). DE genes in the Δ*camA* mutant (sporulation phases) were found to have a significant overlap with DE genes in conditions that favour the production of biofilm; $P<10^{-9}$; the statistical analysis was performed using one-tailed hypergeometric tests implemented in SuperExactTest, Bonferroni adjusted. **f**, Biofilm production, measured by crystal-violet staining absorbance at 570 nm. The differences in biofilm production between Δ*camA* and Δ*camA* N165A could be explained if the Δ*camA* N165A retained some DNA-binding ability and was able to alter the transcription of some genes even in the absence of methylation. Data are mean ± s.d. of $n=3$ independent biological replicates, with each strain assayed in quadruplicate in each experiment. $^*P<0.05$, $^{**}P<10^{-2}$, $^{***}P<10^{-3}$. The statistical analysis was performed using two-way ANOVA with Dunnett's post hoc test.

To identify the regulatory stage of sporulation that CamA-mediated DNA methylation specifically impacts, we used quantitative PCR with reverse transcription (RT–qPCR) to analyse the expression of genes encoding Spo0A, the sporulation-specific sigma factors[32,36] and genes in their individual regulons[32,36,37]. Consistent with our RNA-seq analyses, Spo0A regulon genes—*spo0A, sigF* and *sigE*[32,37]—were expressed at similar levels between WT and Δ*camA* cells at both 9 h and 11 h, implying that the Δ*camA* mutant activates Spo0A at levels similar to the WT. By contrast, the σ^F and σ^E regulon genes, *spoIIQ* and *spoIVA*[32,38], respectively, were underexpressed in Δ*camA* compared with the WT cells (Fig. 5c). Reduced levels of SpoIIQ and SpoIVA were observed in Δ*camA* cells by western blot, confirming the transcriptional analyses (Extended Data Fig. 9c). On the basis of the hierarchical organization of the sporulation regulatory cascade, σ^F activation seems to be the earliest sporulation stage that is affected by CamA. This conclusion is supported by our morphological analyses, because fewer Δ*camA* cells progress to engulfment (a process that requires both σ^F and σ^E activation[39]) than WT cells (Fig. 2c), whereas similar numbers of Δ*camA* and WT cells initiate sporulation. Indeed, similar levels of Spo0A activation are observed in WT and Δ*camA* (Fig. 5c), and the small subset of Spo0A regulon genes that are underexpressed in Δ*camA* cells could be dually regulated by Spo0A and σ^F. For example, *spoIIR*[36]—which encodes a signalling protein required for σ^E activation—is activated by both σ^F and Spo0A[32,35].



In vivo effects of the *camA* mutation. To test whether the sporulation defect of Δ*camA* impacts the infection or transmission of *C. difficile*, we analysed the effect of the Δ*camA* mutation in an established mouse model of infection. Groups of mice (6 males and 6 females) were inoculated by oral gavage with spores of the three genotypes: WT, Δ*camA* and Δ*camA*-C. No mortality was observed at the given doses of *C. difficile* spores, as expected. Faecal samples were collected every 24 h for 7 d. All three *C. difficile* strains reached comparable levels in faeces at days 1 and 2 after inoculation, indicating that they germinate and establish colonization with equal efficiency (Fig. 6a). As expected, colony-forming unit (CFU) levels decreased steadily from day 2 after inoculation to day 7. However, the Δ*camA* mutant showed CFU levels that were 10–100 times lower than those observed in the WT and complemented strains throughout this time frame. The level of bacteria declined to near the limit of detection in the faeces 6 d after inoculation for the MTase mutant, whereas the WT and complemented strains remained detectable at days 6 and 7.

To test whether the loss of CamA leads to defects in virulence, we compared *C. difficile* Δ*camA* and WT in a hamster model of infection. Clindamycin-treated golden Syrian hamsters are highly susceptible to the effects of the *C. difficile* toxins and, therefore, represent a model of acute disease. Groups of 6 hamsters were inoculated by oral gavage with spores of the WT, Δ*camA* and Δ*camA*-C strains. These strains of *C. difficile* elicited diarrhoeal symptoms and weight loss in hamsters, and we observed no difference in the survival times of hamsters after inoculation (Fig. 6b). This result is consistent with the observation that the WT, Δ*camA* and Δ*camA*-C strains exhibit no differences in toxin gene expression (Supplementary Table 8a) and produce comparable levels of TcdA in vitro (Extended Data Fig. 9d). Together, these data indicate that CAAAAA methylation by CamA does not influence toxin-mediated aspects of *C. difficile* pathogenesis but, instead, impacts the ability of *C. difficile* to persist within the host intestinal tract.

Additional functional effects of the *camA* mutation. Considering the high conservation of *camA* across *C. difficile* genomes, we examined whether some additional phenotypes could be effected by the inactivation of *camA*. While analysing images of sporulating *C. difficile*, we noticed that Δ*camA*-mutant cells appeared to be shorter on average than the WT. To test this possibility, we measured the lengths of WT and Δ*camA* cells during broth culture and sporulation and found that Δ*camA* cells were around 15% shorter than the WT cells (Fig. 6c,d) even though no difference in growth was observed (Extended Data Fig. 5b). Interestingly, genes that encode putative cell-wall remodelling enzymes were overexpressed in the Δ*camA* mutant cells relative to the WT during growth in broth culture (Extended Data Fig. 9e).

We next performed an overlap analysis between the list of DE genes from our RNA-seq data (WT versus Δ*camA* mutant; four different time points) and those from published studies focusing on the colonization and infection by this pathogen (Supplementary Table 8f). First, DE genes in the Δ*camA* mutant (sporulation phases) had a significant overlap with DE genes in conditions that favour the production of biofilm on a solid substrate[40] (Fig. 6e). Motivated by this significant overlap, we performed crystal-violet staining assays of the biomass of adherent biofilm, and consistently observed that the Δ*camA* mutant produced more biofilm than the WT cells (Fig. 6f). These results suggest that methylation inhibits the expression of genes that promote biofilm formation. Second, significant overlaps were found when comparing with genes that are DE during infection in different mice gut microbiome compositions[41] (Extended Data Fig. 10a, Supplementary Table 8f). Finally, significant overlaps were found when comparing with DE genes obtained from mice gut isolates at increasing time points after infection[42] (Extended Data Fig. 10b, Supplementary Table 8f). Collectively, these integrative analyses provide further evidence that DNA methylation events by CamA may directly and/or indirectly affect the expression of multiple genes involved in the in vivo colonization and biofilm formation of *C. difficile* and inspire future studies to elucidate the mechanisms that underlie the functional roles of CAAAAA methylation in the pathogenicity of *C. difficile*.

Discussion

C. difficile is responsible for one of the most common hospital-acquired infections and is classified by the US Centers for Disease Control and Prevention as an urgent healthcare risk associated with substantial morbidity and mortality[43]. As *C. difficile* infection is spread by bacterial spores that are found within faeces, extensive research has been devoted to better understand the genome of this important pathogen and its sporulation machinery. To address these common goals, we performed a comprehensive characterization of the DNA methylation landscape across a diverse collection of clinical isolates. During our analysis, we identified a 6mA MTase (*camA*) that is conserved across all of the isolates (and in another ~300 published *C. difficile* genomes) that share a common methylation motif (CAAAAA). Inactivation of the gene encoding this MTase resulted in a sporulation defect in vitro (Fig. 2). Infection studies using a mouse model indicate a role for CamA in the persistence of *C. difficile* in the intestinal tract. As enumeration of *C. difficile* recovered from faeces of the infected animals reflects the number of *C. difficile* spores in the gut, the reduced burden of Δ*camA* in mice might be due to the mutant's defect in sporulation (Fig. 6a), as the ability to form spores was previously shown to be important for persistence[44]. The comparable virulence between Δ*camA* cells and the WT in the hamster model suggests that DNA methylation does not impact toxin-mediated disease. However, owing to the pleiotropic nature of the MTase it remains possible that multiple factors contribute to the more pronounced effect that is observed in the mouse model.

The highly conserved nature of *camA* and its flanking genes across *C. difficile* genomes suggests that additional phenotypes may be regulated by CamA beyond sporulation. Consistent with this, we found CAAAAA sites are over-represented in a set of regions enriched in genes with functions linked to sporulation, motility and membrane transport. Further supporting a broader regulatory network of CamA, the loss of CamA reduces cell length and results in a statistically significant overlap between the transcriptional signatures identified in our study (WT versus Δ*camA* mutant cells) and those of others observed during the in vivo colonization and biofilm formation (Fig. 6e, Extended Data Fig. 10a,b).

The fact that *camA* is a solitary MTase gene without a cognate restriction gene further supports the view that widespread methylation in bacteria has functional importance beyond the role that is attributed to R–M systems. Previously, the most extensively characterized 6mA MTase was Dam-targeting GATC in *Escherichia coli*. Dam has multiple important functions and is essential in some pathogens[12]. However, because Dam is conserved in the large diversity of Enterobacteria, it was not considered to be promising drug target. By contrast, the uniqueness of *camA* in all of the *C. difficile* genomes and in just a few *Clostridiales* makes it a promising drug target that may inhibit *C. difficile* in a much more specific manner, which is particularly relevant because gut dysbiosis potentiates *C. difficile* infection[45,46]. Furthermore, as this MTase does not seem to impact the general fitness of *C. difficile*[23], a drug that specifically targets it may be developed with a lower chance for resistance.

Considering the large number of genes that were DE in the Δ*camA* mutant, the functional impact of CAAAAA methylation is probably mediated by multiple genes that are either directly regulated by DNA methylation or indirectly regulated by a transcriptional cascade. Mechanistically, DNA methylation can either activate or repress a gene depending on other DNA-binding proteins that compete with DNA MTases[7,8,12,47]; therefore, the

competition between TFs and MTases may form an epigenetic switch to turn a gene on and off.

With more than 2,200 bacterial methylomes published to date, it is becoming increasingly evident that epigenetic regulation of gene expression is highly prevalent across bacterial species. Despite the exciting prospects for studying epigenetic regulation, our ability to comprehensively analyse bacterial epigenomes is limited by a bottleneck in integratively characterizing methylation events, methylation motifs, transcriptomic data and functional genomic data. In this regard, this study provides a comprehensive comparative analysis of a large collection of a single bacterial species, as well as a detailed roadmap that can be used by the scientific community to leverage the current status quo of epigenetic analyses.

Methods

***C. difficile* isolates and culture.** We obtained 36 clonal *C. difficile* isolates from infected faecal samples using protocols that were developed in the ongoing Pathogen Surveillance Program at Mount Sinai Hospital (Supplementary Table 1). Furthermore, 9 fully sequenced and assembled *C. difficile* genomes were retrieved from GenBank RefSeq (ftp://ftp.ncbi.nih.gov/genomes, accessed November 2016; Supplementary Table 1). Raw sequencing data from global and UK collections comprising 291 *C. difficile* 027/BI/NAPI genomes were used[3] (Supplementary Table 4). *C. difficile*-positive stool samples were frozen at −80 °C before analysis. All of the stool samples underwent culture for *C. difficile* using an ethanol-shock culture method[48]. In brief, approximately 80 mg of solid stool (50 µl liquid stool samples) was added to 0.5 ml of 70% ethanol wash and the sample was mixed using a vortex and incubated at room temperature for 20 min. A loopful was then cultured onto *C. difficile* selective agar (CDSA, Becton Dickinson) and the plates were incubated anaerobically at 37 °C for up to 72 h. A single colony was subcultured onto a trypticase soy agar plate with 5% defibrinated sheep blood (TSA II, Becton Dickinson) and incubated anaerobically at 37 °C for 48 h. Colonies that had the *C. difficile* odour and showed fluorescence under illumination with ultraviolet light were then obtained and confirmed by matrix-assisted laser desorption/ionization on a Brucker biotyper. For long-term storage, individual colonies were emulsified in tryptic soy broth containing 15% glycerol and stored at −80 °C.

SMRT-seq. Primers were annealed to size-selected (>8 kb) SMRTbell templates with the full-length libraries (80 °C for 2 min 30 s followed by decreasing the temperature to 25 °C by 3 °C min^{-1}). The polymerase–template complex was then bound to P6 enzyme using a ratio of 10:1 polymerase to SMRTbell at 0.5 nM for 4 h at 30 °C and then held at 4 °C until ready for magnetic-bead loading, before sequencing. The magnetic-bead-loading step was performed at 4 °C for 60 min according to the manufacturer's guidelines. The magnetic-bead-loaded polymerase-bound SMRTbell libraries were placed onto the RSII machine at a sequencing concentration of 125–175 pM and configured for a 240 min continuous sequencing run.

De novo genome assembly and motif discovery. The RS_HGAP3 protocol was used for de novo genome assembly, followed by the use of custom scripts for genome finishing and annotation (https://github.com/powerpak/pathogendb-pipeline). RS_Modification_and_Motif_Analysis.1 was used for de novo methylation motif discovery. A custom script was used to examine each motif to ensure its reliable methylation states. In brief, variations in a putative motif were examined by comparing the distribution of the interpulse duration ratio of each variation with non-methylated motifs.

The presence and conservation of *camA* in *C. difficile* isolates. To investigate the pervasive role and conservation of *camA*, we searched for its presence in a global and UK collection of *C. difficile* 027/BI/NAP1 ($n = 291$)[3] genomes (Supplementary Table 4). For this, SRA Illumina reads were converted to FASTQ files using fastq-dump v.2.8.0 and subsequently mapped to the *C. difficile* 630 reference genome using Bowtie2 v.2.2.9 (ref. [49]) in paired-end mode. The resulting SAM files were converted to BAM format (unmapped reads and PCR duplicates were removed) and sorted using SAMTOOLS v.1.9 (ref. [50]). To assess coverage, sequence depths were computed using the genomeCov function of BEDTOOLS v.2.26.0 (ref. [51]) for each strand separately. Variant sites were called from the aligned reads using the mpileup and bcftools tools in SAMTOOLS.

Identification of defence systems. Identification of R–M systems was performed as previously described[52]. In brief, curated reference protein sequences of types I, II, IIC and III R–M systems, and type IV REases were downloaded from the dataset 'gold standards' of REBASE[53] (accessed November 2016). All-against-all searches were performed for REase and MTase standard protein sequences retrieved from REBASE using BLASTP v.2.5.0+ (default settings, $e < 10^{-3}$). The resulting e values were log-transformed and used for clustering into protein families by Markov Clustering v.14-137 (ref. [54]). Each protein family was aligned with MAFFT v.7.305b (ref. [55]) using the E-INS-i option, 1,000 cycles of iterative refinement and offset 0. The alignments were visualized in SEAVIEW v.4.6.1 (ref. [56]) and manually trimmed to remove poorly aligned regions at the extremities. Hidden Markov model (HMM) profiles were then built from each multiple sequence alignment (available at https://github.com/pedrocas81) using the hmmbuild program from the HMMER v.3.0 suite[57] (default parameters). Types I, II and III R–M systems were identified by searching genes encoding the MTase and REase components at less than five genes apart. CRISPR repeats were identified using the CRISPR Recognition Tool (CRT) v.1.2 (ref. [58]) with default parameters. For the CRISPR spacer homology search, hits with at least 80% identity were considered to be positive. For *cas* gene identification, we obtained Cas protein family HMMs from the TIGRFAM database[59] v.15.0 and PFAM families annotated as Cas families (https://ftp.ncbi.nih.gov/pub/wolf/_suppl/CRISPRclass/crisprPro.html). In total we collected 129 known Cas protein families (98 TIGRFAMS and 31 PFAMs), which were used for similarity searching. Genes pertaining to abortive infection systems were searched with the PFAM profiles PF07751, PF08843 and PF14253 (accessed January 2018). Bacteriophage Exclusion (BREX) systems were searched using PFAM profiles for the core genes *pglZ* (PF08665) and *brxC* or *pglY* (PF10923) and specific PFAM profiles for each BREX type as indicated previously[60]. DISARM systems were identified using the PFAM signature domains (PF09369, PF00271 and PF13091) that belong to the core gene triplet characteristic of this system[61]. To search for prokaryotic Argonaute (pAgo) genes, we built a dedicated HMM profile on the basis of a list of 90 Ago-PIWI proteins[62]. Searches for the ensemble of newly found antiphage systems were performed using a previously published list of PFAM profiles[63]. Type II toxin–antitoxin systems were detected using the TAFinder tool[64] with default parameters. Matches of CRISPR spacers were performed against well-known *C. difficile* phages as follows: five siphophages (ϕCD111 (NC_028905.1), ϕCD146 (NC_028958.1), ϕCD38-2 (NC_015568.1), ϕCD6356 (NC_015262.1) and ϕCD211 (NC_029048.2)); five small-tail myophages (ϕMMP04 (NC_019422.1), ϕCD506 (NC_028838.1), ϕCDHM11 (NC_029001.1), ϕCD481-1 (NC_028951.1) and ϕCDHM13 (NC_029116.1)); five medium-tail myophages (ϕMMP03 (NC_028959.1), ϕCDMH1 (NC_024144.1), ϕC2 (NC_009231.1), ϕCD119 (NC_007917.1)) and ϕCDHM19 (NC_028996.1)); and four long-tail myophages (ϕCD27 (NC_011398.1), ϕMMP02 (NC_019421.1), ϕCD505 (NC_028764.1) and ϕMMP01 (NC_028883.1)).

Identification and classification of prophages, conjugative/mobilizable elements and integrons. Prophages were detected with Phage Finder v.2.1 (ref. [65]) using strict mode and PHASTER[66] using the default settings. We took the common hits obtained by both programs, as well as those very few cases (~10% of the hit list) that corresponded to complete prophages predicted by just one of the programs. All elements that were either smaller than 18 kb or lacking matches to core phage proteins (such as terminase, capsid, head and tail proteins) were removed. Integrons were searched using IntegronFinder[67] with the default settings. The identification of genes encoding the functions related to conjugation in ICEs was performed as previously described[68]. In brief, an element was considered to be conjugative when it contained the following components of the conjugative system: a VirB4/TraU ATPase, a relaxase, a coupling ATPase (T4CP) and a minimum number of mating pair formation (MPF) type-specific genes—two for types MPF_{FA} and MPF_{FATA}, or three for the others (types F, T and G). IMEs were identified by the fact that they encode relaxases but lack a complete conjugative transfer system, which is encoded in *trans* by another mobile element. Delimitation of ICEs and IMEs was performed considering flanking core genes as upper bounds for their extremities.

Phylogenetic analyses. The reference phylogenetic tree of *C. difficile* was built from the concatenated alignment of protein families of the core-genome using MUSCLE[69] v.3.8.31 (default parameters). As the DNA sequences provide more phylogenetic signal than protein sequences at this evolutionary distance, we back-translated the alignments to DNA. Poorly aligned regions were removed using BMGE[70] v.1.12. The tree was computed using RAxML[71] v.8.00 under the GTR model and a gamma correction (GAMMA) for variable evolutionary rates. We performed 100 bootstraps on the concatenated alignment to assess the robustness of the topology of the tree.

Identification of the core and pan-genome. The *C. difficile* core genome was built using a previously published methodology[72]. In brief, a preliminary list of orthologues was identified as reciprocal best hits using end-gap-free global alignment between the proteome of a pivot (*C. difficile* 630) and each of the other strain's proteomes. Hits with less than 80% similarity in amino-acid sequence or more than 20% difference in protein length were discarded. This list of orthologues was then refined for every pairwise comparison using information on the conservation of the gene neighbourhood. Positional orthologues were defined as bi-directional best hits adjacent to at least four other pairs of bi-directional best hits within a neighbourhood of ten genes (five upstream and five downstream). The core genome of each clade was defined as the intersection of pairwise lists of positional orthologues. The pan-genome was built using the complete gene repertoire of *C. difficile*. We determined a preliminary list of putative homologous proteins between pairs of genomes by searching for sequence similarity between

each pair of proteins using BLASTP (default parameters). We then used the *e* values ($<10^{-4}$) of the BLASTP output to cluster these proteins using SILIX[73] v.1.2.11. We set the parameters of SILIX such that two proteins were clustered in the same family if the alignment had at least 80% identity and covered more than 80% of the smallest protein (options –I 0.8 and –r 0.8). Core- and pan-genome accumulation curves were built using a dedicated R script. Regression analysis for the pan-genome was performed as described previously[74] by the Heap's power law $n = k \cdot N^{\gamma}$ where n is the pan-genome family size, N is the number of genomes and k, γ ($\alpha = 1 - \delta$) are specific fitting constants. For $\alpha > 1$ ($\delta < 0$) the pan-genome is considered to be closed, that is, sampling more genomes will not affect its size. For $\alpha < 1$ ($0 < \delta < 1$) the pan-genome remains open and the addition of more genomes will increase its size.

Inference of homologous recombination. We inferred homologous recombination on the multiple alignments of the core-genome of *C. difficile* (ordered locally collinear blocks (LCBs) obtained by progressiveMauve were used) using ClonalFrameML[75] v.10.7.5 and Geneconv[76] v.1.81a. The first used a predefined tree (that is, the species tree), default priors $R/\theta = 10^{-1}$ (ratio of recombination and mutation rates), $1/\delta = 10^{-3}$ (inverse of the mean length of recombination events) and $\nu = 10^{-1}$ (average distance between events), and 100 pseudo-bootstrap replicates, as previously suggested[75]. Mean patristic branch lengths were computed with the R package ape[77] v.3.3, and transition/transversion ratios were computed using the R package PopGenome[78] v.2.1.6. The priors estimated by this mode were used as initialization values to rerun ClonalFrameML under the 'per-branch model' mode with a branch dispersion parameter of 0.1. The relative effect of recombination to mutation (r/m) was calculated as $r/m = R/\theta \times \delta \times \nu$. Geneconv was used with options /w123 to initialize the program's internal random number generator and –Skip_indels, which ignores all of the sites with missing data.

Reconstruction of the evolution of gene repertoires. We assessed the dynamics of gene family repertoires using Count[79] (downloaded in January 2018). This program uses birth–death models to identify the rates of gene deletion, duplication and loss in each branch of a phylogenetic tree. We used presence/absence pan-genome matrix and the phylogenetic birth-and-death model of Count to evaluate the most likely scenario for the evolution of a given gene family on the clade's tree. Rates were computed using default parameters, assuming a Poisson distribution for the family size at the tree root and uniform duplication rates. We computed 100 rounds of rate optimization with a convergence threshold of 10^{-3}. After optimization of the branch-specific parameters of the model, we performed ancestral reconstructions by computing the branch-specific posterior probabilities of evolutionary events, and inferred the gains in the terminal branches of the tree. The posterior probability matrix was converted into a binary matrix of presence/absence of horizontal gene transfer genes using a threshold probability of gain higher than 0.2 at the terminal branches. To control for the effects of the choices made in the definition of our model, we computed the gain/loss scenarios using the Wagner parsimony (same parameters, relative penalty of gain with respect to loss of 1). The horizontal gene-transfer events inferred by maximum likelihood and those obtained under Wagner's parsimony were highly correlated (Spearman's $\rho = 0.96$, $P < 10^{-4}$).

Strain construction and growth conditions. The 630Δ*erm*Δ*pyrE* parental strain was used for *pyrE*-based allele-coupled exchange (ACE[80]). A list of *C. difficile* and *E. coli* strains are provided in Supplementary Table 5a. *C. difficile* strains were grown from frozen stocks on brain–heart-infusion-supplemented (BHIS)[81] medium plates supplemented with taurocholate (TA, 0.1% w/v; 1.9 mM), kanamycin (50 μg ml^{-1}) and cefoxitin (8 μg ml^{-1}) as needed. For ACE, *C. difficile* defined medium (CDDM)[82] was supplemented with 5-fluoroorotic acid at 2 mg ml^{-1} and uracil at 5 μg ml^{-1}. Cultures were grown at 37 °C under anaerobic conditions in a gas mixture containing 85% N_2, 5% CO_2 and 10% H_2. The growth curves were performed in BHIS media with gentle shaking. *E. coli* strains were grown at 37 °C with shaking at 225 r.p.m. in Luria–Bertani broth. The medium was supplemented with chloramphenicol (20 μg ml^{-1}) and ampicillin (50 μg ml^{-1}) as needed.

***E. coli* strain construction.** The primers used in this manuscript are provided in Supplementary Table 5b. *C. difficile* 630 genomic DNA was used as the template. To clone the pMTL-YN3-Δ*camA* construct, primer pairs 2332/2334 and 2333/2335 were used to amplify the region 662 bp upstream and 226 bp downstream of *CD630_27580*, respectively. The resulting PCR products were cloned into pMTL-YN3 using Gibson assembly[83]. This construct encodes a *CD630_27580* deletion in which the first 14 codons are linked to the last 139 codons with an intervening stop codon between the 5′ and 3′ end of the gene to avoid production of the last 139 amino acids of CamA. To clone the *camA* complementation constructs, primer pair 2286/2287 was used to amplify *camA* and 163 bp of its upstream region. The resulting PCR product was recombined into pMTL-YN1C by Gibson assembly. The N165A complementation construct was cloned in a similar manner, except that the primer pairs consisted of 2286/2532 and 2531/2287. The plasmids were transformed into *E. coli* DH5α, and the resulting plasmids were confirmed by sequencing and then transformed into HB101/pRK24 for conjugations.

***C. difficile* strain construction.** ACE was used to construct 630Δ*erm*Δ*pyrE*Δ*camA* using uracil and 5-fluoroorotic acid to select for plasmid excision as previously described[84]. The flanking primer pair 2274/2279 was used to screen for the *camA* deletion as shown in Extended Data Fig. 5a (the primers are provided in Supplementary Table 5b). Colonies that appeared to harbour gene deletions were validated by performing an internal PCR using a primer (2288) that binds within the region deleted and a primer (2279) that binds to the region flanking the deletion. Two independent clones from the allelic exchange were phenotypically characterized. The *camA* complementation strains were constructed as previously described using CDDM plates to select for restoration of the *pyrE* locus by recombination[84]. Two independent clones from each complementation strain were phenotypically characterized.

Cell length measurements. Cells were grown to mid-log and stationary phase in BHIS broth or sporulation was induced as described below for three biological replicates. Cells were imaged using phase-contrast microscopy on a Zeiss Axioskop with a ×100/1.3 NA Zeiss Plan Neofluar objective at each time point. Cell length was calculated using the MicrobeJ plugin for Fiji/ImageJ[85]. Image thresholding was performed using the local default method in MicrobeJ/Fiji to account for variations in background. Cell detection parameters were optimized (area, 0–20 μm^2; length, 1 μm maximum; width, 0.5–1 μm) and contours were generated using an interpolated rod-shaped method. Cell length data were exported from MicrobeJ and analysed using Prism 8 (GraphPad).

Sporulation. *C. difficile* strains were inoculated from glycerol stocks overnight onto BHIS–TA plates. Liquid BHIS cultures were inoculated from colonies arising on these plates. The cultures were grown to early stationary phase, back-diluted 1:50 into BHIS, grown until they reached an OD_{600} of between 0.35 and 0.75, and then 120 μl of this culture was spread onto 70:30 (70% SMC media and 30% BHIS media) plates (40 ml). Sporulating cultures were collected into phosphate-buffered saline (PBS), the sample was pelleted and sporulation levels were visualized by phase-contrast microscopy as previously described[25].

Fluorescence microscopy. Fluorescence microscopy was performed on sporulating cultures using Hoechst 33342 (Molecular Probes; 15 μg ml^{-1}) and FM4-64 (Invitrogen; 1 μg ml^{-1}) to stain nucleoid and membrane, respectively. Cells were mounted on a 1% agarose in PBS pad. The images were acquired using a Nikon 80i upright epifluorescence microscope with a Nikon ×60/1.4 NA plan apochromat phase-contrast objective in 12-bit format using Nikon NIS elements software. The images were processed using Adobe Photoshop CC for adjustment of brightness, contrast levels and pseudocolouring.

Spore purification. Sporulation was induced on four 70:30 plates for 48–65 h for each strain tested as described above, and spores were purified as previously described[86]. In brief, sporulating cultures were scraped up, washed repeatedly in ice-cold water, incubated overnight in water on ice, treated with DNase I (New England Biolabs) at 37 °C for 45–60 min and then purified on a density gradient (Histodenz, Sigma Aldrich). Spores were resuspended in 600 μl water for final storage at 4 °C. Spore purity was assessed using phase-contrast microscopy (>95% pure) and the OD_{600} was measured. Spore purification yields were determined from three independent spore preparations. Statistical significance was determined using one-way ANOVA with Tukey's test.

Heat-resistance assay. Heat-resistant spore formation was measured in sporulating *C. difficile* cultures after 20–24 h as previously described[25]. The H_{RES} efficiency represents the average ratio of heat-resistant CFUs to total CFUs for a given strain relative to the average ratio determined for the WT. H_{RES} was determined on the basis of the average H_{RES} values for a given strain in three biological replicates. Statistical significance was determined using one-way ANOVA with Tukey's test.

Germination assay. Germination assays were performed as previously described[34]. Spores (OD_{600} of 0.35, corresponding to ~1×10^7) were resuspended in 100 μl of water, and 10 μl of this mixture was removed for tenfold serial dilutions in PBS. The dilutions were plated on BHIS–TA, and colonies arising from germinated spores were enumerated after 18–21 h. Germination efficiencies were calculated by averaging the CFUs produced by spores for a given strain relative to the number produced by the WT spores for three biological replicates. Statistical significance was determined by performing one-way ANOVA on natural log-transformed data with Tukey's test. The data were transformed because the use of independent spore preparations resulted in a non-normal distribution. Regardless, no statistical significance in germination efficiency was observed for the mutant and its complements.

Spore chloroform resistance. Spores (OD_{600} of 0.75, corresponding to around 2×10^7 spores) were resuspended in 190 μl water. Then, 90 μl of the resuspension was added to tubes containing either 10 μl of water or chloroform for 15 min, after which 10 μl of the sample was serially diluted in PBS and plated on BHIS–TA as described previously[86,87].

CAAAAA motif abundance and exceptionality. We evaluated the exceptionality of the CAAAAA motif using R′MES[26] v.3.1.0. This tool computes scores of exceptionality for *k*-mers of length *l*, by comparing observed and expected counts under Markov models that take sequence composition into consideration. R′MES outputs scores of exceptionality, which are—by definition—obtained from *P* values through the standard one-to-one probit transformation. Analysis of motif abundance was performed using a previously developed framework[27] involving an MSR of genomic signals. We created a binary genomic signal for motif content, which was 1 at motif positions and 0 otherwise. We used 50 length scales. Pruning parameter values were set to default and the *P*-value threshold was set to 10^{-6}.

Whole-genome multiple alignment and classification of CAAAAA positions. Whole-genome multiple alignment of 37 genomes (36 *C. difficile* isolates and *C. difficile* 630) was produced using the progressiveMauve program[88] v.2.4.0 with default parameters. As progressiveMauve does not rely on annotations to guide the alignment, we first used the Mauve Contig Mover[89] to reorder and reorient draft genome contigs according to the reference genome of *C. difficile* 630. A core alignment was built after filtering and concatenating LCBs of at least 50 bp using the stripSubsetLCBs script (http://darlinglab.org/mauve/snapshots/2015/2015-01-09/linux-x64/). The lower value chosen for LCB size accounts for the specific aim of maximizing the number of orthologous motifs detected. The XMFA output format of Mauve was converted to VCF format using dedicated scripts, and VCFtools[90] was used to parse positional variants (SNPs and indels). Orthologous occurrences of the CAAAAA motif were defined if an exact match to the motif was present in each of the 37 genomes (conserved orthologous positions) or if at least one motif (and a maximum of $n-1$, with n being the number of genomes) contained positional polymorphisms (maximum of two SNPs or indels per motif; variable orthologous positions). Non-orthologous occurrences of CAAAAA were obtained from the whole-genome alignment before the extraction of LCBs and correspond to those situations in which the CAAAAA motif was absent in at least one genome. Typically, these correspond to regions containing MGEs or unaligned repetitive regions.

Identification of TFBSs and TSSs. Identification of TFBSs was performed by retrieving *C. difficile* 630 regulatory sites in FASTA format from the RegPrecise database[91] (accessed July 2017). These were converted to position-specific scoring matrices (PSSMs) using in-house-developed scripts. This led to a total of 21 PSSMs pertaining to 14 distinct TF families (Supplementary Table 7b). Matches between these matrices and *C. difficile* genomes were performed using MAST[92] (default settings). MAST output was filtered on the basis of *P* value. Hits with $P<10^{-9}$ were considered to be positive, whereas hits of $P>10^{-5}$ were considered to be negative. Hits with intermediate *P* values were only considered to be positive if the *P* value of the hit divided by the *P* value of the worst positive hit was lower than 100. For the CcpA, LexA, NrdR and CodY TFs (which have shorter binding sites), hits with $P<10^{-8}$ were considered to be positive. TSSs were predicted with Parseq[93] using the 'fast' speed option from multiple RNA-seq datasets (see below). Transcription and breakpoint probabilities were computed using a background expression level threshold of 0.1 and a score penalty of 0.05. We retained only high-confidence 5′ breakpoint hits, located at a maximum distance of 200 bp from the nearest start codon. A ±5 bp window around the TSS was considered if only one single predicted value was obtained; otherwise, we considered an interval that was delimited by the minimum and maximum values predicted by Parseq.

RNA processing. For analyses of sporulating cell transcriptomes, RNA was extracted from three biological replicates of WT and Δ*camA* growing on 70:30 sporulation media after 9 h and 10.5 h of growth using the FastRNA Pro Blue Kit (MP Biomedical) and the FastPrep-24 automated homogenizer (MP Biomedical), similar to previous research[32]. For analyses of the mid-log and early stationary phase cultures, overnight cultures of WT and Δ*camA* in BHIS were back-diluted 1:50 into three biological replicates of 30 ml of BHIS in 125 ml Erlenmeyer flasks. The cultures were grown until mid-log phase ($OD_{600}=0.5–0.6$) and early stationary phase ($OD_{600}=1.3–1.4$). RNA was collected from 15 ml and 10 ml of the same cultures for the mid-log and early stationary-phase cultures, respectively. Contaminating genomic DNA was depleted using three successive DNase treatments, with the final treatment being on-column using the Qiagen RNeasy kit. The samples were tested for genomic DNA contamination using qPCR for 16S rRNA and the *sleC* gene. DNase-treated RNA (15 µg) was enriched for mRNA using the Ribo-Zero Magnetic Kit (Epicentre) for the broth-grown cultures. Ribosomal RNA was depleted from RNA that was collected from sporulating cultures using the Ambion MICROBExpress Bacterial mRNA Enrichment Kit (Thermo Fisher) because Ribo-Zero kits were temporarily discontinued. The quality of total RNA was validated using an Agilent 2100 Bioanalyzer. Samples for RT–qPCR analyses were collected in triplicate from a separate set of three biological replicates that was grown identically to the cultures used for RNA-seq analyses. The RNA was processed similarly except that mRNA enrichment was performed using a MICROBExpress Bacterial mRNA Enrichment Kit, and the DNase-treated RNA samples for RT–qPCR analyses were tested for genomic DNA contamination using qPCR for *rpoB*.

RNA-seq, read alignment and differential-expression analysis. Purified RNA was extracted from three biological replicates of sporulating (9 h and 10.5 h) and exponential and stationary growth cultures of *C. difficile* 630Δ*erm* and *C. difficile* 630Δ*erm*Δ*camA*, DNase-treated, ribosomal RNA-depleted and converted to cDNA as previously described[32]. RNA-seq was performed using a HiSeq 2500, yielding 29.4 ± 4.5 million (mean ± s.d.) 100 bp single-end reads per sample (exponential and stationary growth time points) and 26.9 ± 4.3 million (mean ± s.d.) 150 bp paired-end reads per sample (sporulation time points). Read quality was checked using FastQC v.0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). We used Trimmomatic[94] v.0.39 to remove adapters and low-quality reads (parameters: PE, -phred33, ILLUMINACLIP:<adapters.fa>:2:30:10:8:True, SLIDINGWINDOW:4:15, LEADING:20 TRAILING:20, MINLEN:50). Subsequently, rRNA sequences were filtered from the dataset using SortMeRNA[95] v.2.1 on the basis of the SILVA 16S and 23S rRNA databases[96] and the Rfam 5S rRNA database[97]. The resulting non-rRNA reads were mapped to the *C. difficile* 630 reference genome using BWA-MEM v.0.7.17-r1198 (ref. [98]). The resulting BAM files were sorted and indexed using SAMTOOLS, and read assignment was performed using featureCounts[99] v.1.6.4 (excluding multi-mapping and multi-overlapping reads). A gene was included for differential-expression analysis if it had more than one count in all of the samples. Normalization and differential-expression testing were performed using the Bioconductor package DESeq2 v.1.18.1 (ref. [100]). DE genes were defined as genes with an FDR-corrected $P<0.05$ and $|\log_2[FC]|>1$. Functional classification of genes was performed using the DAVID online database (https://david.ncifcrf.gov)[28]. GO annotation terms with a gene count of at least 5 and $P<0.05$ (one-tailed Fisher's exact test, FDR corrected) were considered to be significant. The reproducibility of DAVID's functional classification was tested with Blast2GO[101] v.5.2 and Panther[102] v.14. In brief, for Blast2GO, we ran BLASTX searches of the *C. difficile* 630 genome against the entire GenBank bacterial protein database (as of September 2018). The output, in XML format, was loaded into Blast2GO, and mapping, annotation and enrichment analysis was performed as indicated (http://docs.blast2go.com/user-manual/quick-start/). For Panther, we downloaded the most recent HMM library (ftp.pantherdb.org/hmm_scoring/13.1/PANTHER13.1_hmmscoring.tgz) and annotated our *C. difficile* 630 protein set with pantherScore2.1.pl. Both input and background gene lists were formatted to the Panther Generic Mapping File type as described at http://www.pantherdb.org. To assess the significance of the intersection between multiple datasets of DE genes (typically observed during *C. difficile* colonization and infection), we collected gene-expression data from in vivo and in vitro studies[40–42], in which key factors for gut colonization (such as time after infection, antibiotic exposure and spatial structure (planktonic and biofilm growth)) were tested. DE genes were called using the same conditions as described above. Statistical analyses and graphical representation of multiset intersections were performed using the R package *SuperExactTest*[103].

RT–qPCR. Transcript levels were determined from cDNA templates that were prepared from the three biological replicates described above. Gene-specific primer pairs are provided in Supplementary Table 5b. RT–qPCR was performed as described previously[33], except that we used iTaq Universal SYBR Green supermix (BioRad), 50 nM of gene specific primers and a Mx3005P qPCR system (Stratagene) in a total volume of 25 µl. The following cycling conditions were used: 95 °C for 2 min, followed by 40 cycles of 95 °C for 15 s and 60 °C for 1 min. Transcript levels were normalized to the housekeeping gene *rpoB* using the standard curve method.

Western blots. *Sporulation protein analyses.* Sporulation was induced as indicated, and samples were collected and processed for immunoblotting as described previously[86]. Total protein in each sample was quantified using the Pierce 660 nm protein assay with ionic detergent compatibility reagent (Thermo Fisher), and 5 µg of protein was loaded for each sample. σ^F, σ^E and Spo0A were resolved using 15% SDS–polyacrylamide gel electrophoresis (SDS–PAGE) gels, whereas SpoIIQ and SpoIVA were resolved using 12% SDS–PAGE gels. Proteins were transferred to polyvinylidene difluoride membranes, which were subsequently probed with rabbit (anti-σ^F, anti-σ^E and anti-SpoIIQ) and mouse (anti-Spo0A and anti-SpoIVA) polyclonal primary antibodies, and anti-rabbit IR800 and anti-mouse IR680 secondary antibodies (LI-COR). Blots were imaged using an LiCor Odyssey CLx imaging system. The results shown are representative of analyses of two biological replicates.

Toxin analyses. Overnight cultures of *C. difficile* were diluted 1:50 in TY medium and incubated at 37 °C for 24 h. Cells were collected using centrifugation, suspended in SDS–PAGE buffer and boiled for 10 min. The samples were then run on 4–20% Mini-PROTEAN TGX Precast Protein Gels (Bio Rad) and transferred to a nitrocellulose membrane. TcdA was detected as described previously using mouse anti-TcdA primary antibodies (Novus Biologicals) and goat anti-mouse IgG conjugated with IR800 (Thermo Fisher)[104].

Animal infection studies. All of the animal experiments was performed under the guidance of veterinarians and trained animal technicians within the University of North Carolina Division of Comparative Medicine. Animal experiments were

performed with prior approval from the UNC Institutional Animal Care and Use Committee. Animals considered to be moribund as defined in the protocols were euthanized by CO_2 asphyxiation followed by a secondary physical method in accordance with the Panel on Euthanasia of the American Veterinary Medical Association. The University complies with state and federal Animal Welfare Acts, the standards and policies of the Public Health Service.

Mouse model. The parental *C. difficile* strain 630Δ*erm*, the MTase mutant 630Δ*erm*Δ*camA* and the MTase complemented strain were evaluated in an antibiotic-treated mouse model as previously described[105,106]. Groups of 8-to-10-week old female and male C57BL/6 mice (*Mus musculus*; Charles River Laboratories) were administered a cocktail of antibiotics (kanamycin (400 μg ml^{-1}), gentamicin (35 μg ml^{-1}), colistin (850 U ml^{-1}), vancomycin (45 μg ml^{-1}) and metronidazole (215 μg ml^{-1})) in their water ad libitum 7 d before inoculation for 3 d, followed by a single intra-peritoneal dose of clindamycin (10 mg kg^{-1} body weight) 2 d before inoculation. The mice were randomly assigned into groups, with 2 female mice assigned to the mock condition and 6 mice (3 male and 3 female) to each infection condition. The experiment was independently repeated to assess the consistency of the data. The data from the experiments were combined for analysis for a total of 12 mice (6 male and 6 female) in each infection condition. Mice were inoculated with 10^5 spores by oral gavage. Mock-inoculated mice were included as controls. Cage changes were performed every 48 h after inoculation. Faecal samples were collected every 24 h for 7 d after inoculation. Dilutions were plated on BHIS agar containing 0.1% of the germinant TA to enumerate spores as CFU g^{-1} of faeces.

Hamster model. The above strains were tested in Syrian golden hamster strain LVG (*Mesocricetus auratus*; Charles River Laboratories) as described previously[107]. Hamsters were randomly assigned into groups, with 2 assigned to the mock condition and 6 (3 male and 3 female) to each infection condition. Hamsters were administered a single dose of clindamycin (30 mg kg^{-1} body weight) by oral gavage, then inoculated with approximately 5,000 spores of the above strains 5 d later. Hamsters were monitored for weight loss and diarrhoeal symptoms and were considered to be moribund after 15–20% weight loss from their maximum body weight, with or without concurrent diarrhoea.

Biofilm assays. Biofilm assays were performed as previously described[108]. In brief, overnight cultures of *C. difficile* were diluted 1:100 in BHIS supplemented with 1% glucose and 50mM sodium phosphate buffer (pH 7.5) in 24-well polystyrene plates. After 24 h of growth at 37 °C, supernatants were removed, the biofilms were washed once with PBS and then stained for 30 min with 0.1% (w/v) crystal violet. After 30 min, the biofilms were washed again with PBS, and the crystal violet was solubilized with ethanol. Absorbance was read at 570 nm. Three independent experiments were performed, and each strain was assayed in quadruplicate in each experiment.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Genome assemblies and methylation data are available from NCBI under BioProject ID PRJNA448390. RNA-seq data are available under project ID PRJNA445308. Additional data are available from the corresponding authors on reasonable request.

Code availability

Scripts and a tutorial supporting all of the key analyses of this research are publicly available as a package named Bacterial Epigenome Analysis SuiTe (BEAST) at http://github.com/fanglab/.

Received: 14 August 2018; Accepted: 18 October 2019;
Published online: 25 November 2019

References

1. Smits, W. K., Lyras, D., Lacy, D. B., Wilcox, M. H. & Kuijper, E. J. *Clostridium difficile* infection. *Nat. Rev. Dis. Primers* **2**, 16020 (2016).
2. Sebaihia, M. et al. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat. Genet.* **38**, 779–786 (2006).
3. He, M. et al. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat. Genet.* **45**, 109–113 (2013).
4. Herbert, M., O'Keeffe, T. A., Purdy, D., Elmore, M. & Minton, N. P. Gene transfer into *Clostridium difficile* CD630 and characterisation of its methylase genes. *FEMS Microbiol. Lett.* **229**, 103–110 (2003).
5. van Eijk, E. et al. Complete genome sequence of the *Clostridium difficile* laboratory strain 630Δ*erm* reveals differences from strain 630, including translocation of the mobile element CTn*5*. *BMC Genom.* **16**, 31 (2015).
6. Hargreaves, K. R., Thanki, A. M., Jose, B. R., Oggioni, M. R. & Clokie, M. R. Use of single molecule sequencing for comparative genomics of an environmental and a clinical isolate of *Clostridium difficile* ribotype 078. *BMC Genom.* **17**, 1020 (2016).
7. Casadesus, J. & Low, D. Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.* **70**, 830–856 (2006).
8. Low, D. A., Weyand, N. J. & Mahan, M. J. Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. *Infect. Immun.* **69**, 7197–7204 (2001).
9. Cohen, N. R. et al. A role for the bacterial GATC methylome in antibiotic stress survival. *Nat. Genet.* **48**, 581–586 (2016).
10. Manso, A. S. et al. A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nat. Commun.* **5**, 5055 (2014).
11. Atack, J. M. et al. A biphasic epigenetic switch controls immunoevasion, virulence and niche adaptation in non-typeable *Haemophilus influenzae*. *Nat. Commun.* **6**, 7828 (2015).
12. Wion, D. & Casadesus, J. N^6-methyl-adenine: an epigenetic signal for DNA-protein interactions. *Nat. Rev. Microbiol.* **4**, 183–192 (2006).
13. Oliveira, P. H., Touchon, M. & Rocha, E. P. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl Acad. Sci. USA* **113**, 5658–5663 (2016).
14. Flusberg, B. A. et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461–465 (2010).
15. Beaulaurier, J., Schadt, E. E. & Fang, G. Deciphering bacterial epigenomes using modern sequencing technologies. *Nat. Rev. Genet.* **20**, 157–172 (2019).
16. Fang, G. et al. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat. Biotechnol.* **30**, 1232–1239 (2012).
17. Murray, I. A. et al. The methylomes of six bacteria. *Nucleic Acids Res.* **40**, 11450–11462 (2012).
18. Davis, B. M., Chao, M. C. & Waldor, M. K. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. *Curr. Opin. Microbiol.* **16**, 192–198 (2013).
19. Smits, W. K. Hype or hypervirulence: a reflection on problematic *C. difficile* strains. *Virulence* **4**, 592–596 (2013).
20. Roberts, R. J. et al. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* **31**, 1805–1812 (2003).
21. Wust, J., Sullivan, N. M., Hardegger, U. & Wilkins, T. D. Investigation of an outbreak of antibiotic-associated colitis by various typing methods. *J. Clin. Microbiol.* **16**, 1096–1101 (1982).
22. Barra-Carrasco, J. & Paredes-Sabja, D. *Clostridium difficile* spores: a major threat to the hospital environment. *Future Microbiol.* **9**, 475–486 (2014).
23. Dembek, M. et al. High-throughput analysis of gene essentiality and sporulation in *Clostridium difficile*. *mBio* **6**, e02383 (2015).
24. Donnelly, M. L., Fimlaid, K. A. & Shen, A. Characterization of *Clostridium difficile* spores lacking either SpoVAC or dipicolinic acid synthetase. *J. Bacteriol.* **198**, 1694–1707 (2016).
25. Shen, A., Fimlaid, K. A. & Pishdadian, K. Inducing and quantifying *Clostridium difficile* spore formation. *Methods Mol. Biol.* **1476**, 129–142 (2016).
26. Schbath, S. & Hoebeke, M. in *Advances in Genomic Sequence Analysis and Pattern Discovery* Vol. 7 (eds Elnitsk, L. et al.) 25–64 (World Scientific, 2011).
27. Knijnenburg, T. A. et al. Multiscale representation of genomic signals. *Nat. Methods* **11**, 689–694 (2014).
28. Huang, D. W. et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* **35**, W169–W175 (2007).
29. Lim, H. N. & van Oudenaarden, A. A multistep epigenetic switch enables the stable inheritance of DNA methylation states. *Nat. Genet.* **39**, 269–275 (2007).
30. Ardissone, S. et al. Cell cycle constraints and environmental control of local DNA hypomethylation in α-Proteobacteria. *PLoS Genet.* **12**, e1006499 (2016).
31. Cota, I. et al. OxyR-dependent formation of DNA methylation patterns in OpvABOFF and OpvABON cell lineages of *Salmonella enterica*. *Nucleic Acids Res.* **44**, 3595–3609 (2016).
32. Fimlaid, K. A. et al. Global analysis of the sporulation pathway of *Clostridium difficile*. *PLoS Genet.* **9**, e1003660 (2013).
33. Pishdadian, K., Fimlaid, K. A. & Shen, A. SpoIIID-mediated regulation of σK function during *Clostridium difficile* sporulation. *Mol. Microbiol.* **95**, 189–208 (2015).
34. Fimlaid, K. A. & Shen, A. Diverse mechanisms regulate sporulation sigma factor activity in the *Firmicutes*. *Curr. Opin. Microbiol.* **24**, 88–95 (2015).
35. Saujet, L., Pereira, F. C., Henriques, A. O. & Martin-Verstraete, I. The regulatory network controlling spore formation in *Clostridium difficile*. *FEMS Microbiol. Lett.* **358**, 1–10 (2014).

36. Saujet, L. et al. Genome-wide analysis of cell type-specific gene transcription during spore formation in *Clostridium difficile*. *PLoS Genet.* **9**, e1003756 (2013).
37. Rosenbusch, K. E., Bakker, D., Kuijper, E. J. & Smits, W. K. *C. difficile* 630Δerm Spo0A regulates sporulation, but does not contribute to toxin production, by direct high-affinity binding to target DNA. *PLoS ONE* **7**, e48608 (2012).
38. Fimlaid, K. A., Jensen, O., Donnelly, M. L., Siegrist, M. S. & Shen, A. Regulation of *Clostridium difficile* spore formation by the SpoIIQ and SpoIIIA proteins. *PLoS Genet.* **11**, e1005562 (2015).
39. Ribis, J. W., Fimlaid, K. A. & Shen, A. Differential requirements for conserved peptidoglycan remodeling enzymes during *Clostridioides difficile* spore formation. *Mol. Microbiol.* **110**, 370–389 (2018).
40. Maldarelli, G. A. et al. Type IV pili promote early biofilm formation by *Clostridium difficile*. *Pathog. Dis.* **74**, ftw061 (2016).
41. Jenior, M. L., Leslie, J. L., Young, V. B. & Schloss, P. D. *Clostridium difficile* colonizes alternative nutrient niches during infection across distinct murine gut microbiomes. *mSystems* **2**, e00063-17 (2017).
42. Fletcher, J. R., Erwin, S., Lanzas, C. & Theriot, C. M. Shifts in the gut metabolome and *Clostridium difficile* transcriptome throughout colonization and infection in a mouse model. *mSphere* **3**, e00089-18 (2018).
43. Lessa, F. C. et al. Burden of *Clostridium difficile* infection in the United States. *N. Engl. J. Med.* **372**, 825–834 (2015).
44. Deakin, L. J. et al. The *Clostridium difficile spo0A* gene is a persistence and transmission factor. *Infect. Immun.* **80**, 2704–2711 (2012).
45. Lewis, B. B. & Pamer, E. G. Microbiota-based therapies for *Clostridium difficile* and antibiotic-resistant enteric Infections. *Annu. Rev. Microbiol.* **71**, 157–178 (2017).
46. Abt, M. C., McKenney, P. T. & Pamer, E. G. *Clostridium difficile* colitis: pathogenesis and host defence. *Nat. Rev. Microbiol.* **14**, 609–620 (2016).
47. Sanchez-Romero, M. A., Cota, I. & Casadesus, J. DNA methylation in bacteria: from the methyl group to the methylome. *Curr. Opin. Microbiol.* **25**, 9–16 (2015).
48. Griffiths, D. et al. Multilocus sequence typing of *Clostridium difficile*. *J. Clin. Microbiol.* **48**, 770–778 (2010).
49. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
50. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
51. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
52. Oliveira, P. H., Touchon, M. & Rocha, E. P. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **42**, 10618–10631 (2014).
53. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE-a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **43**, D298–D299 (2015).
54. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
55. Katoh, K. & Standley, D. M. MAFFT: iterative refinement and additional methods. *Methods Mol. Biol.* **1079**, 131–146 (2014).
56. Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221–224 (2010).
57. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
58. Bland, C. et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinform.* **8**, 209 (2007).
59. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
60. Goldfarb, T. et al. BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.* **34**, 169–183 (2015).
61. Ofir, G. et al. DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat. Microbiol.* **3**, 90–98 (2018).
62. Makarova, K. S., Wolf, Y. I., van der Oost, J. & Koonin, E. V. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biol. Direct* **4**, 29 (2009).
63. Doron, S. et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).
64. Xie, Y. et al. TADB 2.0: an updated database of bacterial type II toxin-antitoxin loci. *Nucleic Acids Res.* **46**, D749–D753 (2018).
65. Fouts, D. E. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* **34**, 5839–5851 (2006).
66. Arndt, D. et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–W21 (2016).
67. Cury, J., Jove, T., Touchon, M., Neron, B. & Rocha, E. P. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.* **44**, 4539–4550 (2016).
68. Cury, J., Touchon, M. & Rocha, E. P. C. Integrative and conjugative elements and their hosts: composition, distribution and organization. *Nucleic Acids Res.* **45**, 8943–8956 (2017).
69. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **5**, 113 (2004).
70. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
71. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
72. Touchon, M. et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* **5**, e1000344 (2009).
73. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinform.* **12**, 116 (2011).
74. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **11**, 472–477 (2008).
75. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
76. Sawyer, S. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**, 526–538 (1989).
77. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
78. Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**, 1929–1936 (2014).
79. Csuros, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912 (2010).
80. Ng, Y. K. et al. Expanding the repertoire of gene tools for precise manipulation of the *Clostridium difficile* genome: allelic exchange using *pyrE* alleles. *PLoS ONE* **8**, e56051 (2013).
81. Sorg, J. A. & Dineen, S. S. Laboratory maintenance of *Clostridium difficile*. *Curr. Protoc. Microbiol.* **12**, 9A.1.1–9A.1.10 (2009).
82. Cartman, S. T. & Minton, N. P. A mariner-based transposon system for in vivo random mutagenesis of *Clostridium difficile*. *Appl. Environ. Microbiol.* **76**, 1103–1109 (2010).
83. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
84. Donnelly, M. L. et al. A *Clostridium difficile*-specific, gel-forming protein required for optimal spore germination. *mBio* **8**, e02085-16 (2017).
85. Ducret, A., Quardokus, E. M. & Brun, Y. V. MicrobeJ, a tool for high throughput bacterial cell detection and quantitative analysis. *Nat. Microbiol.* **1**, 16077 (2016).
86. Ribis, J. W., Ravichandran, P., Putnam, E. E., Pishdadian, K. & Shen, A. The conserved spore coat protein SpoVM Is largely dispensable in *Clostridium difficile* spore formation. *mSphere* **2**, e00315-17 (2017).
87. Edwards, A. N. et al. Chemical and stress resistances of *Clostridium difficile* spores and vegetative cells. *Front. Microbiol.* **7**, 1698 (2016).
88. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).
89. Rissman, A. I. et al. Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* **25**, 2071–2073 (2009).
90. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
91. Novichkov, P. S. et al. RegPrecise 3.0—a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genom.* **14**, 745 (2013).
92. Bailey, T. L. & Gribskov, M. Combining evidence using *p*-values: application to sequence homology searches. *Bioinformatics* **14**, 48–54 (1998).
93. Mirauta, B., Nicolas, P. & Richard, H. Parseq: reconstruction of microbial transcription landscape from RNA-seq read counts using state-space models. *Bioinformatics* **30**, 1409–1416 (2014).
94. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
95. Kopylova, E., Noe, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
96. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
97. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439–441 (2003).
98. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

99. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
100. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
101. Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
102. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).
103. Wang, M., Zhao, Y. & Zhang, B. Efficient test and visualization of multi-set intersections. *Sci. Rep.* **5**, 16923 (2015).
104. Anjuwon-Foster, B. R., Maldonado-Vazquez, N. & Tamayo, R. Characterization of flagellum and toxin phase variation in *Clostridioides difficile* ribotype 012 isolates. *J. Bacteriol.* **200**, e00056-18 (2018).
105. Chen, X. et al. A mouse model of *Clostridium difficile*-associated disease. *Gastroenterology* **135**, 1984–1992 (2008).
106. McKee, R. W., Aleksanyan, N., Garrett, E. M. & Tamayo, R. Type IV pili promote *Clostridium difficile* adherence and persistence in a mouse model of infection. *Infect. Immun.* **86**, e00943-17 (2018).
107. Woods, E. C., Edwards, A. N., Childress, K. O., Jones, J. B. & McBride, S. M. The *C. difficile clnRAB* operon initiates adaptations to the host environment in response to LL-37. *PLoS Pathog.* **14**, e1007153 (2018).
108. Purcell, E. B. et al. A nutrient-regulated cyclic diguanylate phosphodiesterase controls *Clostridium difficile* biofilm and toxin production during stationary phase. *Infect. Immun.* **85**, e00347-17 (2017).
109. Pereira, F. C. et al. The spore differentiation pathway in the enteric pathogen *Clostridium difficile*. *PLoS Genet.* **9**, e1003782 (2013).
110. Serrano, M. et al. A recombination directionality factor controls the cell type-specific activation of σ^K and the fidelity of spore development in *Clostridium difficile*. *PLoS Genet.* **12**, e1006312 (2016).
111. Theriot, C. M. et al. Cefoperazone-treated mice as an experimental platform to assess differential virulence of *Clostridium difficile* strains. *Gut Microbes* **2**, 326–334 (2011).

Acknowledgements

We thank R. J. Roberts (New England Biolabs) for his help with the prediction of R–M systems and orphan MTases in *C. difficile* genomes using REBASE Tools and for providing comments. He was originally an author of this manuscript; however, as a staunch supporter of the open access movement, he will not author a paper that is not open access. We also thank E. P. C. Rocha (Institut Pasteur, Paris, France) for reading the manuscript and for providing comments. The research was primarily funded by R01 GM114472 (to G.F.) from the National Institutes of Health and Icahn Institute for Genomics and Multiscale Biology. The research was also funded by NIH grants R01 AI119145 (to H.v.B and A.B.), R01 AI22232 (to A.S.), R01 AI107029 (to R.T.) and R35 GM131780 (to A.K.A), a Hirschl Research Scholar award from the Irma T. Hirschl/Monique Weill-Caulier Trust (to G.F.), a Pew Scholar in the Biomedical Sciences grant from the Pew Charitable (to A.S.). G.F. is a Nash Family Research Scholar. A.S. holds an Investigators in the Pathogenesis of Infectious Disease Award from the Burroughs Wellcome Fund. J.W.R was supported by an NIH training grant 5T32GM007310-42. The participation of R. J. Roberts in this project was funded by New England Biolabs. This research was also supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Author contributions

G.F. conceived the hypothesis. A.S. and G.F. supervised the project. P.H.O. and G.F. designed the computational methods. P.H.O., R.T., A.S. and G.F. designed the experiments. P.H.O. performed most of the computational analyses and developed most of the scripts that support the analyses. J.W.R. performed the growth curves, microscopy analyses (fluorescence and phase contrast), analyses of cell length and sporulation stage, isolation of some of the RNA and processed it for RT–qPCR studies, and RT–qPCR analyses of sporulation genes. A.S. constructed the deletion and catalytic Δ*camA* mutants, performed complementation, isolated and processed the RNA for several of the RNA analyses, and performed many of the sporulation phenotypic assays. E.M.G. and D.T. performed the animal infection experiment and analysed the data under the supervision of R.T. A.Kim and G.F. performed methylation motif discovery and refinement. O.S. and E.A.M. performed RT–qPCR controls for RNA-seq analyses. O.S., E.A.M., G.D., M.L.-S., C.B., N.E.Z., D.R.A., I.O., G.P., F.W., C.H., S.H., R.S., H.v.B. and A.S. contributed to the other experiments. G.D., I.O. and R.S. designed and conducted SMRT-seq. P.H.O., J.W.R., E.M.G., D.T., A.Kim, O.S., T.P., S.Z., E.A.M., M.T., C.B., S.B., A.K.A., A.B., R.T., E.E.S., R.S., H.v.B., A.Kasarskis, R.T., A.S. and G.F. analysed the data. P.H.O., R.T., A.S. and G.F. wrote the manuscript with additional information inputs from other co-authors.

Competing interests

A.S. has a consultant role for BioVector, a diagnostic start-up. The other authors declare no competing interests.

Additional information

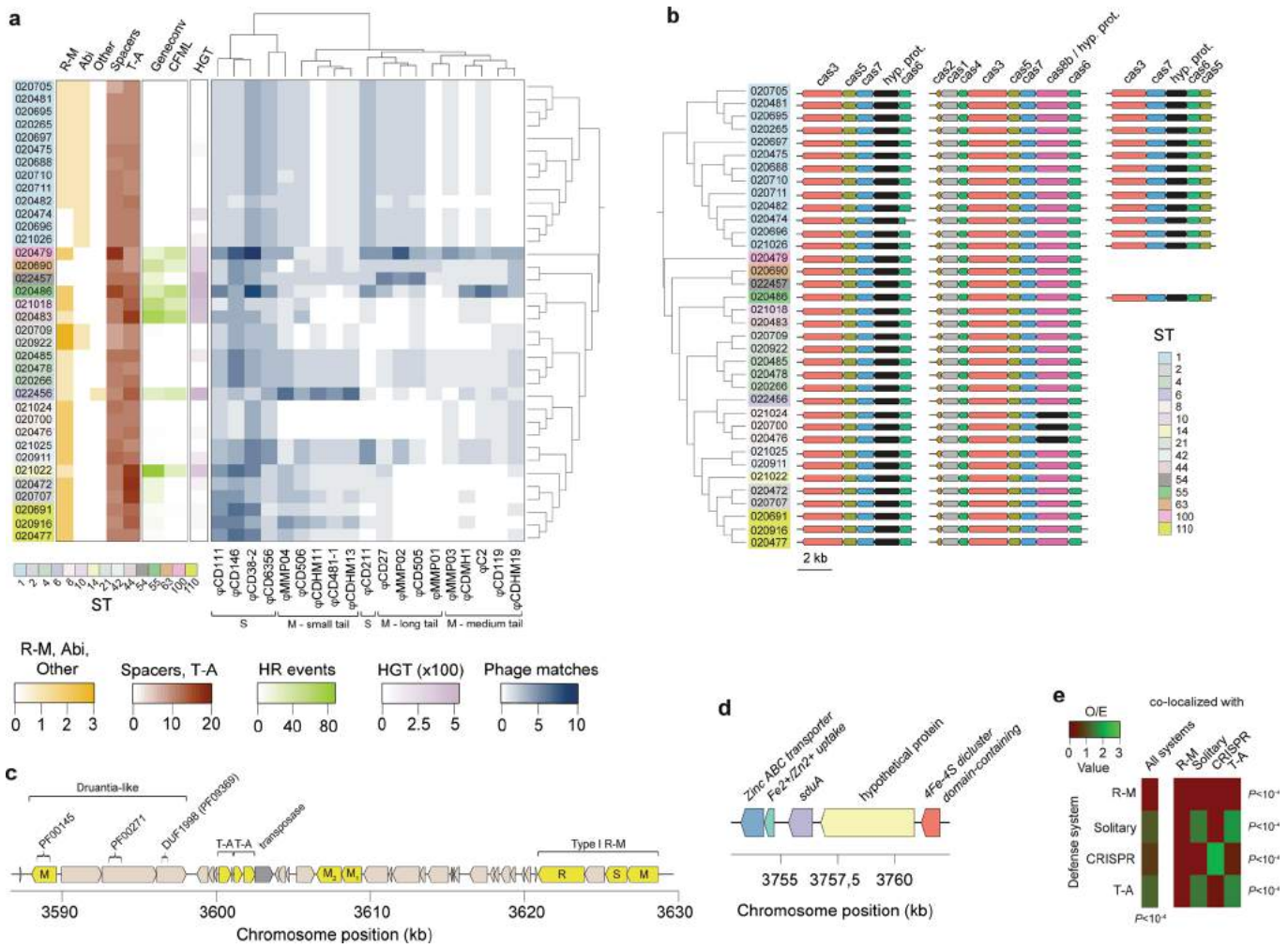
Extended data is available for this paper at https://doi.org/10.1038/s41564-019-0613-4.

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-019-0613-4.

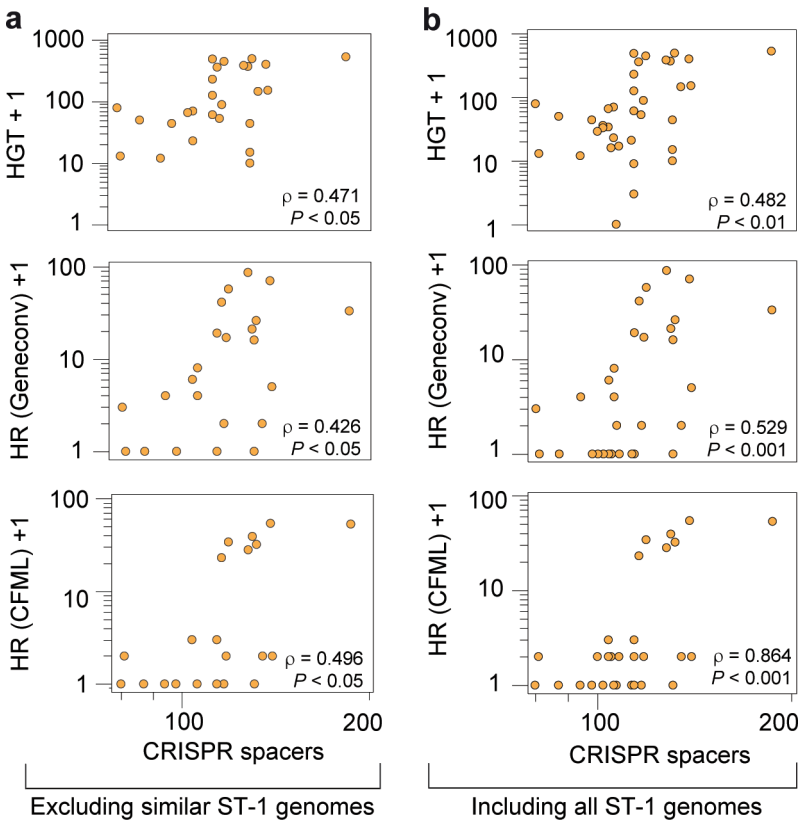
Correspondence and requests for materials should be addressed to A.S. or G.F.

Reprints and permissions information is available at www.nature.com/reprints.

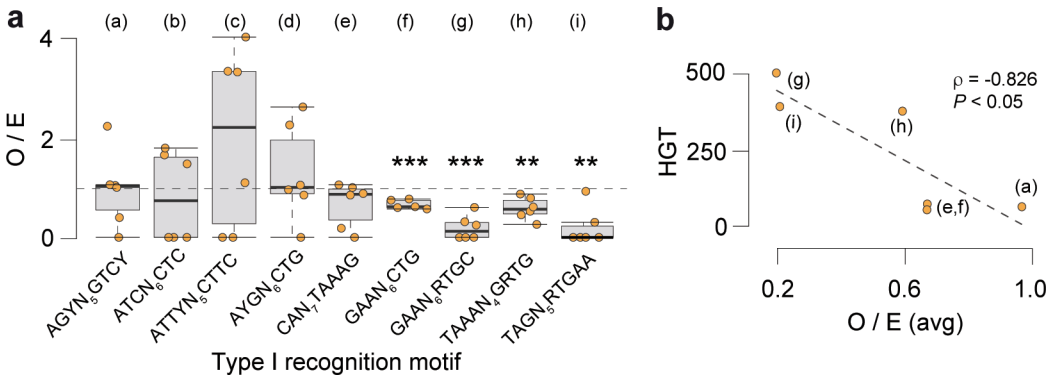
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



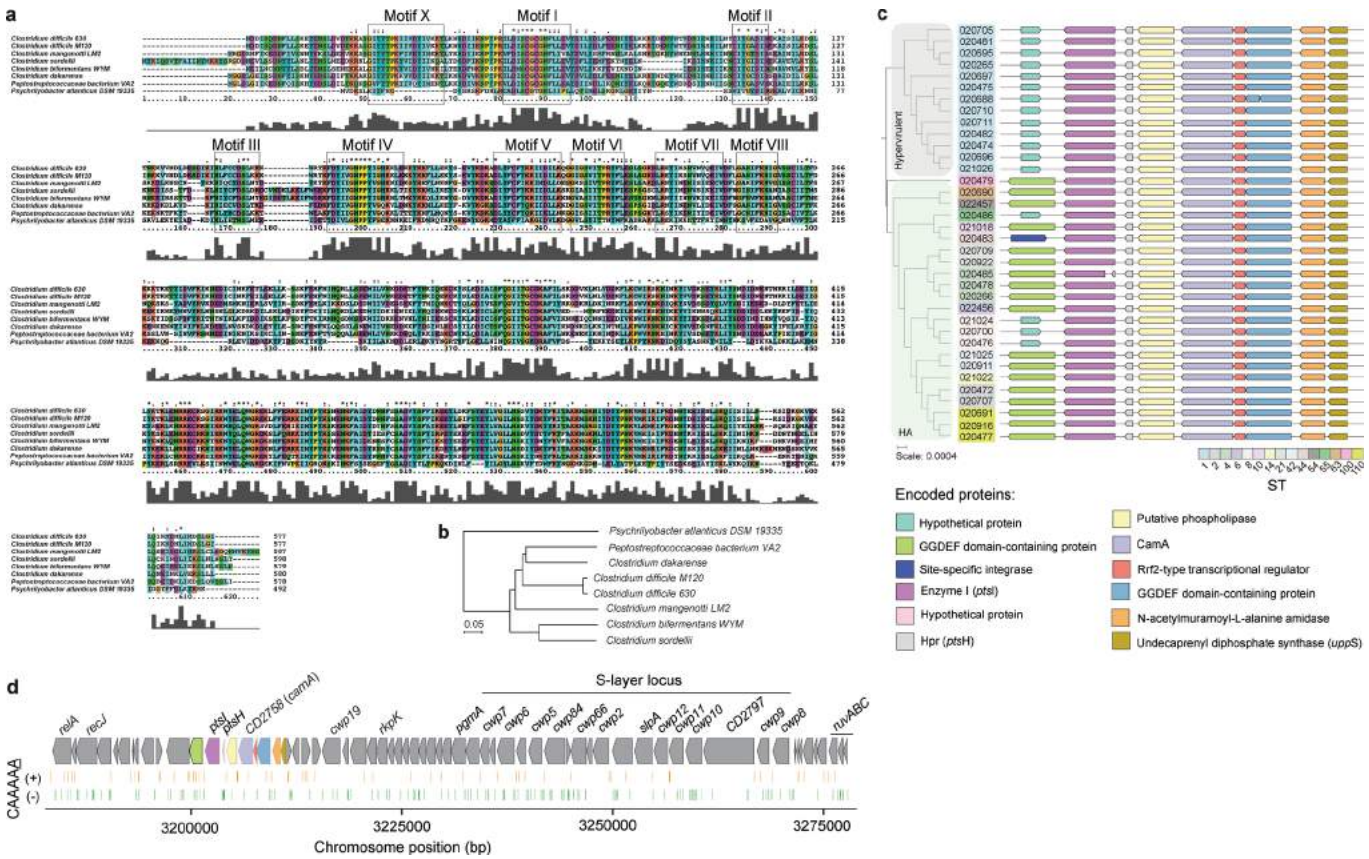
Extended Data Fig. 1 | Multiple defense systems and gene flux control in *C. difficile*. (a) Heatmap aggregate depicts: abundance of defense systems (R-M, abortive infection (Abi), average number of spacers per CRISPR, toxin-antitoxin (T-A), and Shedu systems (other)), homologous recombination (HR) events (given by Geneconv and ClonalFrameML (CFML)), horizontal gene transfer (HGT, given by Wagner parsimony), and number of phage-targeting CRISPR spacers (Supplementary Notes). Phages were clustered according to their family (*Siphoviridae* (S), *Myoviridae* (M)), and tail type. (b) Cas genes detected in *C. difficile*. Apart from the complete Type-IB gene cluster (*cas1-cas8*), we also observed two truncated gene clusters lacking *cas1*, *cas2*, and *cas4*. One of the truncated operons was present across all genomes, while the second was restricted to ST-1 and ST-55. (c) Example of a putative 'defense island' detected in CD_020472 harboring: a Druantia-like system, two T-A systems, two solitary MTases, and one Type I R-M system. The Druantia-like system is similar to the previously reported Type II Druantia systems[63] in the sense that a PF00271 helicase conserved C-terminal domain and DUF1998 (PF09369) are associated with a nearby cytosine methylase. However, it lacks a PF00270 DEAx box helicase. (d) Genomic context of the *sduA* gene in CD_22456 pertaining to the newly identified Shedu defense system. The gene is located in an integrative conjugative element (ICE) (Supplementary Table 2d). (e) Observed/expected (O/E) ratios for co-localized defense systems (maximum of 10 genes apart). Only the most abundant systems were included in the analysis. Expected values were obtained by multiplying the total number of defense systems by the fraction of co-localized defense systems. *P* values correspond to the Chi-square test.



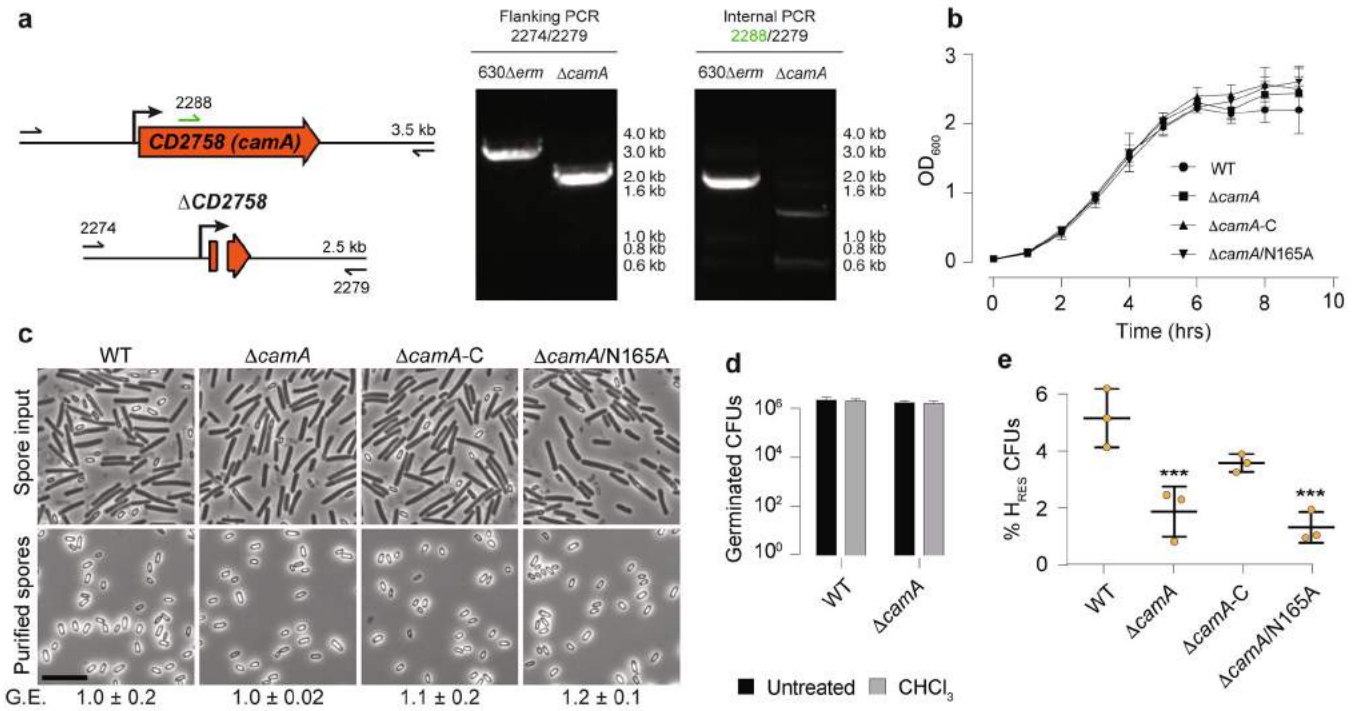
Extended Data Fig. 2 | Relation between gene flux and CRISPR spacer content. (a) Association between genetic flux (horizontal gene transfer (HGT) and homologous recombination (HR, computed using both ClonalFrameML (CFML) and Geneconv)) and number of CRISPR spacers. The latter were used as proxy of their activity. Data was plotted after excluding very similar ST-1 genomes. The criteria to remove these genomes were based on similarities in R-M content, and gene flux, that is, all ST-1 genomes but CD_020475, CD_020474, CD_021026 were removed ($n = 26$). (b) Same as (a) but considering the complete genome dataset ($n = 36$). Spearman's rank correlation coefficients (ρ) and associated P values (two-sided) are shown in each graph.



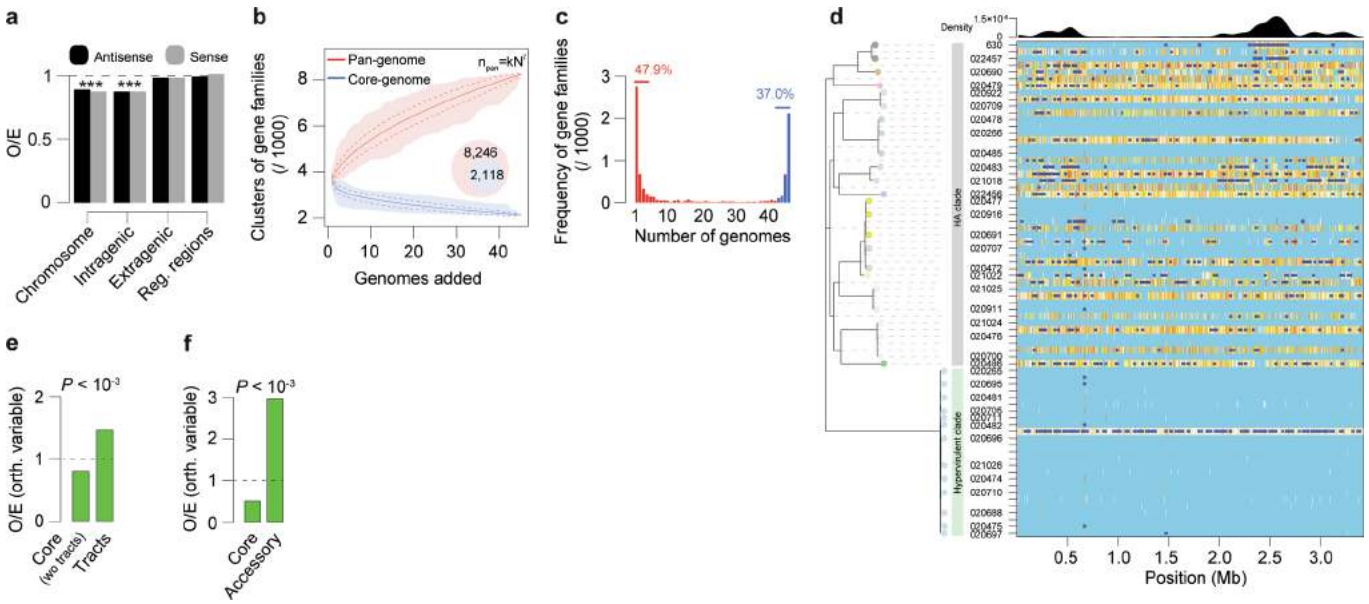
Extended Data Fig. 3 | Interplay between Type I R-M systems and gene flux in *C. difficile*. (a) Observed/expected (O/E) ratios for Type I target recognition motifs in *Clostridioides* phage genomes. 6 phage genomes representative of *Siphoviridae* and *Myoviridae* families and tail types were analyzed (ϕCD111, ϕCDHM11, ϕMMP01, ϕMMP04, ϕC2, ϕCD38). O/E values were obtained with R'MES using Markov chain models that take into consideration oligonucleotide composition. For each motif, we tested if the median value of the O/E ratio in phage genomes was significantly different from 1. In box plots, the middle line indicates the median value, boxes are 25th and 75th quartiles, and whiskers indicate 1.5 times the interquartile range. ***$P < 10^{-3}$; **$P < 10^{-2}$ (one-sided one-sample t-test). (b) Relation between HGT and O/E ratio for Type I target recognition motifs. For those *C. difficile* genomes harboring a single Type I R-M system (that is, without the confounding effect of multiple systems), we computed the average values of HGT, and plotted these values against the average O/E ratio for the corresponding target recognition motif in phage genomes. This was only possible for the $n = 6$ motifs indicated in brackets. The spearman's rank correlation coefficient (ρ) and associated P value (two-sided) is shown.



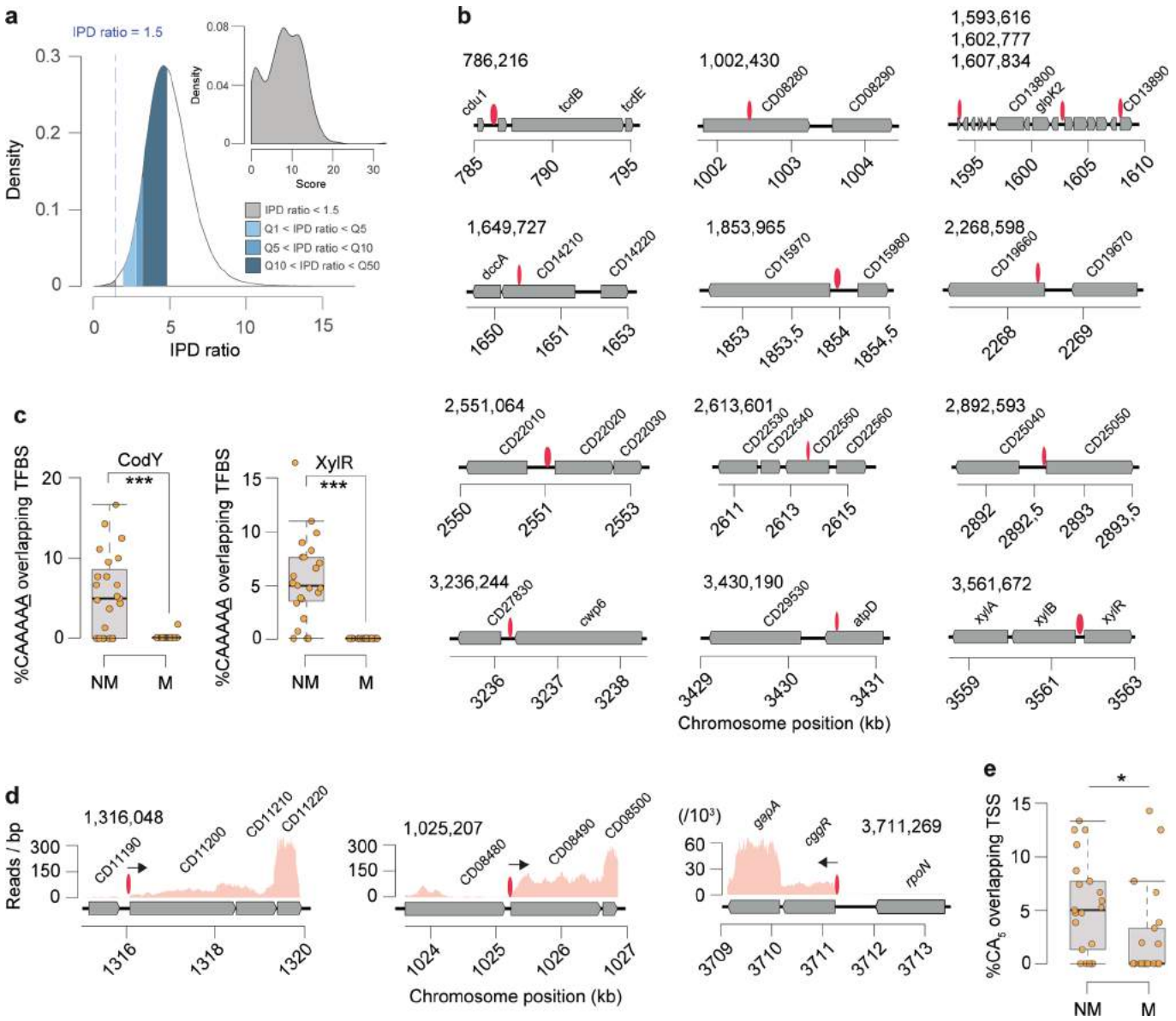
Extended Data Fig. 4 | Genomic context and conservation of *camA*. (a) CamA protein alignment among *Clostridiales* (*C. mangenotti* LM2 (587 aa, 56% identity), *C. sordellii* (598 aa, 53% identity), *C. bifermentans* WYM (579 aa, 53% identity), *C. dakarense sp. nov* (580 aa, 63% identity), *Peptostreptococcaceae bacterium* VA2) and *Fusobacteriales* (*Psychrilyobacter atlanticus* DSM 19335) using ClustalX. The nine conserved motifs (I-VIII and X) typically found in MTases are highlighted. (b) Phylogenetic tree obtained from the MTase alignment. (c) Phylogenetic tree of the 36 *C. difficile* strains colored by clade (hypervirulent, human/animal (HA) associated) and MLST sequence type (ST). Shown is the genomic context of *camA* across the entire dataset. (d) Expanded view of the region shown in Fig. 1f. The example shown (including coordinates) refers to the reference genome of *C. difficile* 630. + and – signs correspond to the sense and antisense strands respectively. Vertical bars correspond to the distribution of the CAAAAA motif.



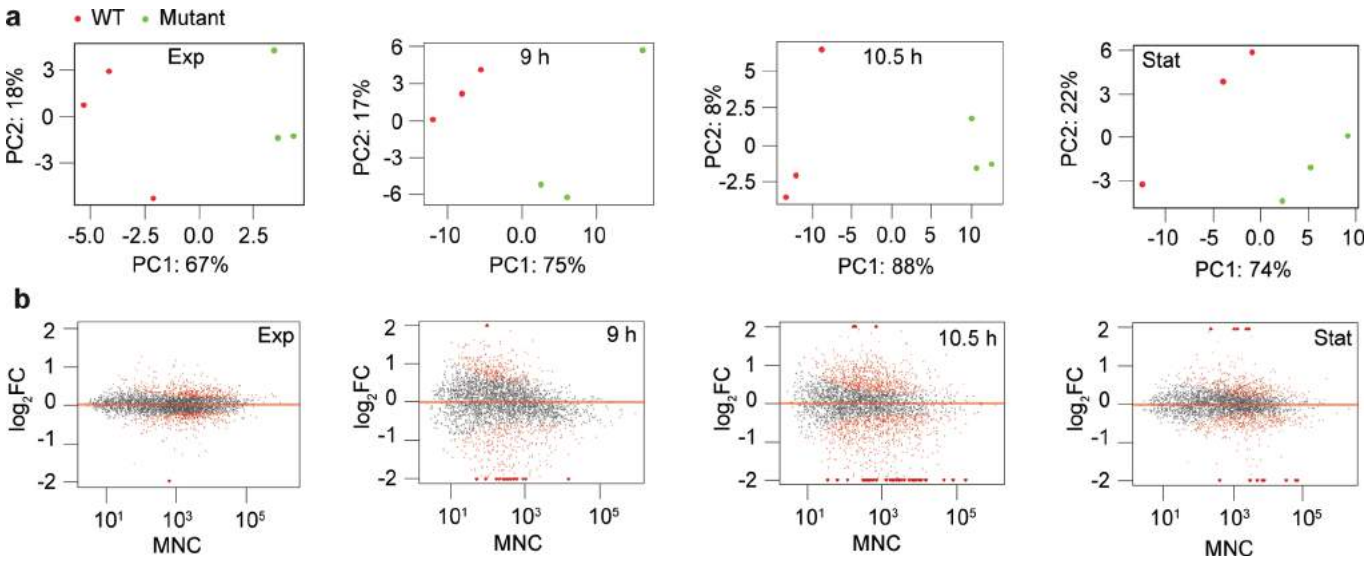
Extended Data Fig. 5 | Δ*camA* construction, purified spore analyses, broth culture growth, and sporulation kinetics. (a) PCR to distinguish between wild-type *camA* and Δ*camA* using flanking primers and primers internal to the deletion. PCRs were performed twice independently. (b) Growth curves comparing wild-type *camA*, Δ*camA*, Δ*camA-C*, and *camA/N165A* cultures grown in BHIS liquid media. Early stationary-phase cultures were diluted to a starting O.D. of 0.05 in BHIS media and growth was measured over 9 h. Each pair genotype / timepoint correspond to mean of $n = 3$ independent biological replicates. Error bars correspond to standard deviation. (c) Phase-contrast microscopy analyses of sporulating culture samples prior to and after spore purification on a density gradient. No gross differences in spore morphology were observed between wild type and the MTase mutant. The germination efficiency (G.E.) of purified spores from the indicated strains is shown below. Scale bar represents 5 μm. Microscopy analyses were performed on three independent spore preparations. (d) Chloroform resistance of purified Δ*camA* spores relative to wild type. Spores were treated with 10 % chloroform for 15 min after which spore viability was measured by plating untreated and chloroform-treated spores on media containing germinant and measuring colony forming units. No significant differences in germination efficiency or chloroform resistance were observed. Data are presented as mean ± standard deviation of four independent biological replicates. (e) Heat-resistance (H_{RES}) efficiencies of sporulating cultures 22 h after sporulation induction were determined relative to wild-type. Data are presented as mean ± standard deviation. Three independent biological replicates per group were used. *** $P < 10^{-3}$, one-way ANOVA with Tukey's test.



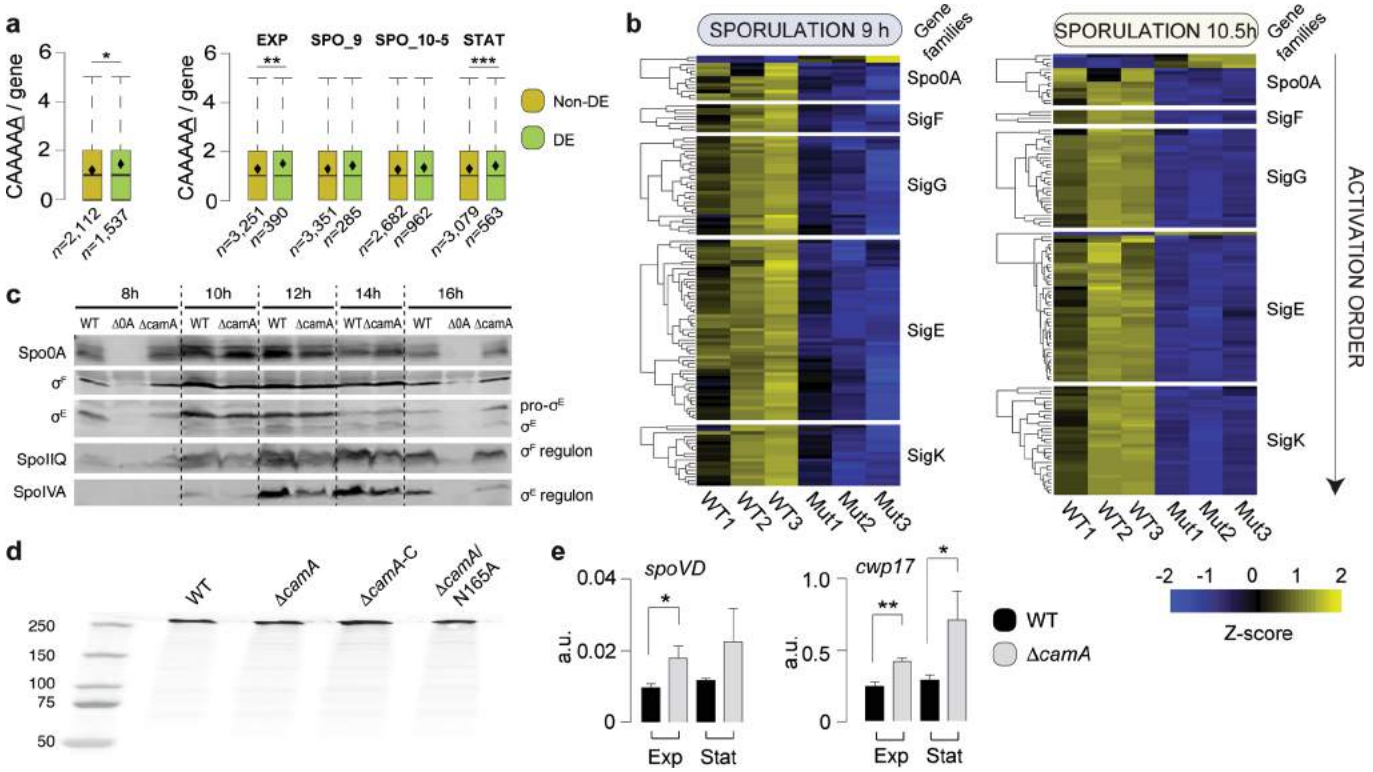
Extended Data Fig. 6 | CAAAAA exceptionality, core- / pan-genome analyses of *C. difficile*, and homologous recombination (HR) landscape. (a) Observed (O) numbers of CAAAAA motifs in the *C. difficile* chromosome ($n=7,824$), intragenic ($n=6,131$), extragenic ($n=1,693$), and regulatory regions ($n=794$, defined as the windows spanning 100 bp upstream the start codon to 50 bp downstream) were compared with expected (E) values computed in random sequences showing the same oligonucleotide composition. The significance of the difference between O/E was evaluated by computing a *P* value based on a Gaussian approximation of motif counts under a Markov model of order 4 (*** $P<10^{-3}$). (b) Core- and pan-genome sizes of *C. difficile*. The pan- and core-genomes were used to perform gene accumulation curves. These curves describe the number of new genes (pan-genome) and genes in common (core-genome) obtained by adding a new genome to a previous set. The procedure was repeated 1,000 times by randomly modifying the order of integration of the $n=45$ genomes in the analysis. Solid lines correspond to the average number of gene families obtained across all permutations, dashed lines indicate standard deviation of the mean, and shaded regions indicate range. The values for the specific constants obtained after Heap's law fitting are 2,887 and 0.271, respectively for the k and γ, thus implying an open pan-genome. (c) Spectrum of frequencies for *C. difficile* gene repertoires. It represents the number of genomes where the families of the pan-genome can be found, from 1 for strain-specific genes to 45 for core-genes. Red indicates accessory genes and blue the genes that are highly persistent in *C. difficile*. (d) Graphical representation of the recombinational events in the core genome of *C. difficile* (inferred by ClonalFrameML). The HA and hypervirulent branches of the tree are depicted in colors. Substitutions are represented by vertical lines and recombination events by dark blue horizontal bars. Light blue vertical lines represent the absence of substitutions, and white lines refer to non-homoplasic substitutions. All other colors represent homoplasic substitutions, with increases in homoplasy associated with increases in the degree of redness (from white to red). (e) O/E ratios of orthologous variable CAAAAA motifs (compared to orthologous conserved) in the core-genome (excluding recombination tracts) ($n=770$) and recombination tracts ($n=325$), or (f) core ($n=1,095$) and accessory genome ($n=1,415$). *P* values correspond to the Chi-square test.



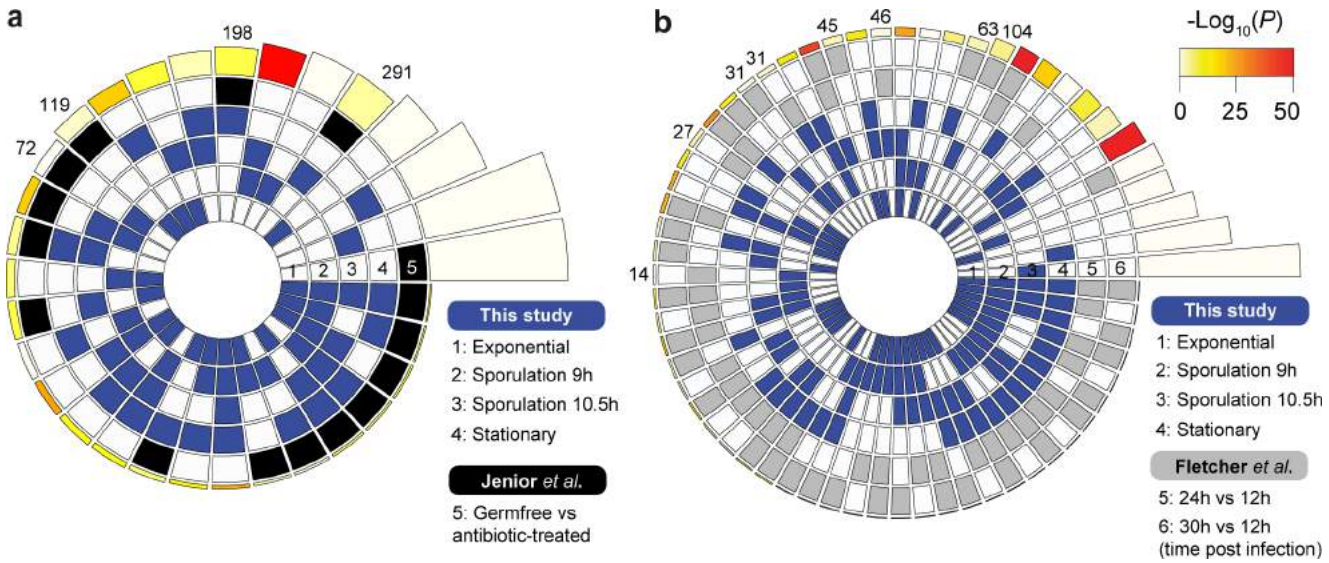
Extended Data Fig. 7 | Non-methylated CAAAAA motif sites overlapping TFBSs and TSSs. (a) Interpulse duration ratio (ipdR) density distribution of the terminal adenine of CAAAAA. Motifs were considered as non-methylated if the terminal adenine had IPD ratios < 1.5 (stippled line), coverage > 20×, and methylation scores < 20 (gray distribution). Also shown for comparison are the sections delimited by quantiles (Q) 1, 5, 10, and 50. (b) Additional examples of highly conserved non-methylated CAAAAA motif sites (red ovals) and corresponding genetic context. Positions indicated above the graph correspond to the non-methylated base. (c) %CAAAAA motif sites (non-methylated (NM) and methylated (M)) overlapping CodY and XylR TFBS for each *C. difficile* isolate excluding ST1 genomes ($n = 23$). (d) Additional examples of chromosomal regions for which non-methylated CAAAAA motif sites overlap TSSs (shown as arrows). (e) %CAAAAA motif sites (non-methylated and methylated) overlapping TSSs for each *C. difficile* isolate excluding ST1 genomes ($n = 23$). For box plots the middle line indicates the median value, boxes are 25th and 75th quartiles, and whiskers indicate 1.5 times the interquartile range. * $P < 0.05$, *** $P < 10^{-3}$ (one-sided Mann-Whitney-Wilcoxon rank sum test with continuity correction).



Extended Data Fig. 8 | Principal Component Analysis (PCA) and MA-plots for RNA-seq data. (a) PCA performed using DESeq2 rlog-normalized RNA-seq data ($n = 3$ biological replicates for each genotype). (b) MA-plots showing the variation of fold change with mean normalized counts (MNC). Number of genes represented: 3,532 (Exp), 3,426 (9 h), 3,523 (10.5), and 3,510 (Stat). Red-colored points have P values < 0.1 (Wald test, Benjamini-Hochberg adjusted). Points that fall out of the window are plotted as open triangles pointing either up or down.



Extended Data Fig. 9 | DE, gene, and protein expression analyses. (a) Enrichment of the CAAAAA motif in DE genes compared to non-DE ones either globally (left, $n = 3{,}649$ genes) or at each time point studied (right, $n_{EXP} = 3{,}641$, $n_{SPO\_9} = 3{,}636$, $n_{SPO\_10.5} = 3{,}644$, $n_{STAT} = 3{,}642$). For box plots, the middle line indicates the median value, boxes are 25th and 75th quartiles, and whiskers indicate 1.5 times the interquartile range. * $P < 0.05$, ** $P < 10^{-2}$, *** $P < 10^{-3}$ (one-sided Mann-Whitney-Wilcoxon rank sum test with continuity correction). (b) Time-course change in the expression of genes under the control of the specific sigma factors (σ^F, σ^E, σ^G, and σ^K) and master transcriptional activator Spo0A at both 9 and 10.5 h after sporulation induction (respectively $n = 121$ and $n = 124$ genes). (c) Representative immunoblot time-course (from $n = 2$ independent biological replicates with similar results) comparing the levels of the early sporulation proteins σ^F, SpoIIQ, σ^E, and SpoIVA in WT and Δ*camA* at 8, 10, 12, 14, and 16 h following induction of sporulation. (d) Western blot for TcdA for each *C. difficile* genotype. (e) RT-qPCR of *spoVD* and *cwp17* genes ($n = 3$ independent biological replicates) of exponential and stationary phase liquid broth cultures. Data is presented as mean ± standard deviation. * $P < 0.05$, ** $P < 10^{-2}$, two-tailed unpaired Student's t-test.



Extended Data Fig. 10 | Overlap between multiple datasets of differentially expressed (DE) genes. Comparisons were performed between DE genes called in this study for each time point (blue-shaded, $n = 1{,}537$) and those obtained from (a) Jenior *et al.* (black-shaded, $n = 971$) and (b) Fletcher *et al.* (gray-shaded, 299). Color intensities of the outermost layer represent the *P* value significance of the intersections (3,896 genes were used as background). The height of the corresponding bars is proportional to the number of common genes in the intersection (indicated at the top of the bars for pairwise comparisons between the different studies). Significant overlaps were found between our DE dataset and either (a) genes DE during infection in different mice gut microbiome compositions ($P < 10^{-6}$, one-tailed hypergeometric test implemented in *SuperExactTest*, Bonferroni adjusted), or (b) DE genes obtained from mice gut isolates at increasing time points after infection ($P < 10^{-4}$, one-tailed hypergeometric test implemented in *SuperExactTest*, Bonferroni adjusted).

nature research

Corresponding author(s): Aimee Shen, Gang Fang

Last updated by author(s): Oct 18, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	Scripts and a tutorial supporting all key analyses of this work are publicly available as a package named Bacterial Epigenome Analysis SuiTe (BEAST) at http://github.com/fanglab/.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Genome assemblies and methylation data are available via NCBI under BioProject ID PRJNA448390. RNA-Seq data are available via under project PRJNA445308. Additional data are available from the corresponding authors upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	36 clonal C. difficile isolates from CDI fecal samples were obtained using protocols developed in an ongoing Pathogen Surveillance Program at Mount Sinai Hospital.
Data exclusions	No data were excluded from the analyses.
Replication	Experiments were performed at least in triplicate unless noted otherwise. All replication attempts were successful.
Randomization	No randomization was used, except for the mice and hamster experiments in which both were randomly assigned.
Blinding	Blinding was not used in our study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	All the antibodies were produced by sending the antigens to CoCalico Biologicals. Full list is shown in the materials and methods section.
Validation	We validated each antibody against a mutant in the publications that were associated with each antibody in the materials and methods section.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	We have used groups of 8- to 10-week old male and female C57BL/6 mice (Mus musculus; Charles River Laboratories). We have also used groups of 5- to 10 week male and female Syrian golden hamsters strain LVG (Mesocricetus auratus; Charles River Laboratories).
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected from the field.
Ethics oversight	All animal experimentation was performed under the guidance of veterinarians and trained animal technicians within the University of North Carolina Division of Comparative Medicine. Animal experiments were performed with prior approval from the UNC Institutional Animal Care and Use Committee. The University complies with state and federal Animal Welfare Acts, the standards and policies of the Public Health Service.

Note that full information on the approval of the study protocol must also be provided in the manuscript.