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

Pedro Oliveira, John Ribis, Elizabeth Garrett, Dominika Trzilova, Alex Kim, Ognjen Sekulovic, Edward Mead, Theodore Pak, Shijia Zhu, Gintaras Deikus, et al.

HHS Public Access

Author manuscript

Nat Microbiol. Author manuscript; available in PMC 2020 May 25.

Published in final edited form as:

Nat Microbiol. 2020 January ; 5(1): 166–180. doi: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], **Robert Sebra**[1], **Harm van Bakel**[1], **Andrew Kasarskis**[1], **Rita Tamayo**[3], **Aimee Shen**[2], **Gang Fang**[1]

[1]Department of Genetics and Genomic Sciences, Institute for Genomics and Multiscale Biology, Mount Sinai School of Medicine, New York, New York, United States of America

[2]Department of Molecular Biology and Microbiology, Tufts University School of Medicine, Boston, Massachusetts, United States of America

[3]Department of Microbiology and Immunology, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, North Carolina, USA

[4]Microbial Evolutionary Genomics, Institut Pasteur, 25–28 rue du Docteur Roux, Paris, 75015, France

[5]CNRS, UMR3525, 25–28 rue du Docteur Roux, Paris, 75015, France

[6]Department of Medicine, Division of Infectious Diseases, Mount Sinai School of Medicine, New York, New York, United States of America

[7]Department of Pharmacological Sciences and Department of Oncological Sciences, Mount Sinai School of Medicine, New York, New York, United States of America

Aimee Shen (aimee.shen@tufts.edu); Gang Fang (gang.fang@mssm.edu).

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 supporting the analyses. J.W.R. performed the growth curves, microscopy analyses (fluorescence and phase-contrast), analyses of cell length and sporulation stage, isolated some of the RNA and processed it for qRT-PCR studies; qRT-PCR 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 analyzed 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 qRT-PCR 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 sequencing. 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.J.B., R.T., E.E.S., R.S., H.v.B., A.Kasarskis., R.T., A.S. and G.F. analyzed 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 Statement

A.S. has a consultant role for BioVector, a diagnostic start-up. We declare that the authors have no other competing interests as defined by Nature Publishing Group, or other interests that might be perceived to influence the results and/or discussion in this article.

Abstract

Clostridioides difficile is a leading cause of health care-associated infections. Although significant 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 performed a comprehensive DNA methylome analysis of *C. difficile* using 36 human isolates and observed great epigenomic diversity. We discovered an orphan DNA methyltransferase with a well-defined specificity whose corresponding gene is highly conserved across our dataset and in all ~300 global *C. difficile* genomes examined. Inactivation of the methyltransferase gene negatively impacted sporulation, a key step in *C. difficile* disease transmission, consistently supported by multi-omics data, genetic experiments, and a mouse colonization model. Further experimental and transcriptomic analysis also suggested 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 critical pathogen. This work also provides a set of methods for comparative epigenomics and integrative analysis, which we expect to be broadly applicable to bacterial epigenomics studies.

Keywords

DNA methylation; SMRT sequencing; biofilm formation; restriction-modification systems

Introduction

Clostridioides (formerly *Clostridium*) *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 significant progress 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: N6-methyladenine (6mA, the most prevalent form representing ~80%), N4-methylcytosine (4mC), and 5-methylcytosine (5mC). Increasing evidence suggests that DNA methylation regulates a number of biological processes such as DNA replication and repair, cell cycle, chromosome segregation and gene expression, among others[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 09/2019) have been mapped, heralding a new era of “bacterial epigenomics”[18].

Herein, we mapped and characterized the DNA methylomes of 36 human *C. difficile* isolates using SMRT-seq and comparative epigenomics. We observed great epigenomic diversity across *C. difficile* isolates, as well as the presence of a highly conserved methyltransferase (MTase). Inactivation of this MTase resulted in a functional impact on sporulation, a key step in *C. difficile* transmission. Further experimental and integrative transcriptomic analysis suggested that epigenetic regulation by DNA methylation also modulates *C. difficile* cell length, host colonization and biofilm formation. These discoveries are expected to stimulate

future investigations along a new epigenetic dimension to characterize and potentially repress medically relevant biological processes in this critical 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 fecal samples of infected patients (Supplementary Table 1). A total of 15 different MLST sequence types (STs) belonging to clades 1 (human and animal, HA1) and 2 (so-called 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 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 (average of 2.6 motifs per genome) (Fig. 1a, Supplementary Table 2a). The large majority of target motifs were of 6mA type, one motif (TAACTG) belonged to the 4mC type, and no confident 5mC motifs were detected (Supplementary Notes). Like most bacterial methylomes, >95% of the 6mA and 4mC motif sites were methylated (Fig. 1b, Supplementary Table 2a).

Genomes pertaining to the same ST tend to have more similar sets of methylation motifs relative to those from different STs. Those belonging to ST-2, ST-8, ST-21, and ST-110 showed the highest motif diversities. One 6mA motif, CAAAAA, was present across all genomes, which led us to hypothesize that 6mA methylation events at this motif, and its corresponding MTase, play 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 motif CAAAAA across all the *C. difficile* isolates, we proceeded to examine the encoded MTases. We identified a total of 139 MTase genes (average of 3.9 per genome) (Fig. 1a, Supplementary Table 2b) representing all the four major types[20], and appearing either in a solitary context or within restriction-modification (R-M) systems (Figs. 1c–e, Supplementary Tables 2b–d). We further found multiple additional defense systems (*e.g.*, abortive infection systems, CRISPR-Cas, toxin-antitoxin), and performed an integrative analysis with R-M systems in relation to host defense and gene flux (Extended Data Figs. 1,2, Supplementary Tables 3a–g), such as that involving phages (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 aa) present across isolates (Fig. 1f, Supplementary Table 2b, Supplementary Notes) and responsible for methylation of the former. This MTase is encoded by *CD2758* in the reference strain *C. difficile* 630[2,21]. Here we have named CD2758 as CamA (***C***. *difficile* **a**denine **m**ethyltransferase **A**). Its ubiquity was not restricted to the 36 isolates, as we were able to retrieve orthologs in a list of ~300 global *C. difficile* isolates from GenBank (Supplementary Table 4). REBASE also showed functional orthologs of *camA* only in very few 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 if *camA* inactivation could reduce spore purification efficiencies in the 630 strain as previously suggested for its homolog in the 027 isolate R20291[23]. We constructed an in-frame deletion in this gene and complemented it with either wild type *camA* or a variant encoding a catalytic site mutation (N165A) of the MTase (Extended Data Fig. 5a; Supplementary Tables 5a, b). We observed that spore purification efficiencies decreased by ~50% in the mutant relative to wild type (Fig. 2a). Complementation of Δ*camA* with the wild type, but not the catalytic mutant, restored spore purification efficiencies to values similar to those observed in wild-type cells (Fig. 2a, Supplementary Table 5c). No differences in growth were observed between wild-type and mutant strains (Extended Data Fig. 5b). Hence, this complementation experiment supports 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 spore purification efficiencies observed in the Δ*camA* mutants could be due to a reduced number of cells inducing 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). Since purified wild-type and Δ*camA* spores had similar levels of chloroform resistance and germinated with similar efficiency (Extended Data Figs. 5d,e), the reduced spore purification efficiencies of the MTase mutants likely reflect a defect in sporulation initiation rather than the sporulation process itself. Accordingly, fewer Δ*camA* cells were observed to be sporulating in phase-contrast microscopy analyses relative to wild type (Fig. 2b).

To gain insight into the sporulation stage affected by loss of CamA, we quantified the number of sporulating cells at different stages of spore assembly (Fig. 2c). While similar numbers of wild-type 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, ~2-fold more Δ*camA* cells were at asymmetric division relative to wild type 11 h after sporulation induction, whereas 50% fewer Δ*camA* cells had completed engulfment. Since similar numbers of sporulating cells were observed between wild type and Δ*camA* at 11 h, Δ*camA*'s sporulation defect appears to arise from fewer cells progressing beyond asymmetric division rather than a defect in sporulation induction.

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

Comparative analysis of CAAAAA sites across *C. difficile* genomes

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

To further characterize CAAAAA 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 either: (1) conserved orthologous (devoid of 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 ones contribute to variations at CAAAAA sites across genomes with subsequent methylation abrogation (Supplementary Table 6f). Such across-genome variation appears to be at least partially fueled by events of homologous recombination (Extended Data Figs. 6b–f, Supplementary Table 6g). Lastly, DAVID gene enrichment analysis found cytoplasm- and motility-related genes to 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 harboring genes related to sporulation and colonization and that orthologous variable CAAAAA positions are enriched in regions harboring 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 (e.g., transcription factors, TFs) as previously described[12,28–30]. Previous bacterial methylome studies analyzing one or few genomes denoted having insufficient statistical power to perform a systematic interrogation of non-methylated motifs sites. 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). Non-methylated motif sites were found dispersed throughout the full length of the *C. difficile* chromosome, yet were overrepresented in orthologous variable and non-orthologous CAAAAA positions (O/E=respectively 1.51 and 1.49) and underrepresented in orthologous conserved CAAAAA positions (O/E 0.84) (all $P < 10^{-4}$; Chi-square 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, while a minor 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 allowed a systematic search for evidence of overlap with 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 (Figs. 4a,b, Extended Data Fig. 7b, Supplementary Tables 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-seq data coverage, we found a similar genome-level enrichment for non-methylated CAAAAA sites (Figs. 4d, e; Extended Data Figs. 7d, e, Supplementary Table 7d; $P < 10^{-3}$, Mann-Whitney-Wilcoxon test,). Hence, these results demonstrate the occurrence of on/off epigenetic switch of CAAAAA sites preferentially overlapping with putative TFBSs and TSSs.

Loss of CAAAAA methylation impacts transcription of multiple gene categories including sporulation

To study the functional significance of methylation at CAAAAA sites, we used RNA-seq to compare the transcriptomes of wild-type *C. difficile* 630 with that of Δ*camA* both in liquid medium (exponential and stationary growth stage) and following sporulation induction (9 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% FDR and $|\log_2 FC| > 1$ (2-fold change in gene expression) (Fig. 5a, Supplementary Tables 8b–d). DE genes in Δ*camA* relative to wild type showed significant enrichment in CAAAAA motif sites compared to 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 loss of CamA reduces spore formation, the transcriptome analyses revealed that 118 and 120 genes previously identified as being induced during sporulation[31,32] were expressed at ≥ 50% lower levels in Δ*camA* cells relative to wild type at 9 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[33,34] (Fig. 5b), with the early-acting sigma factors, σ^F and σ^E, being required for the activity of the later-acting sigma factors, σ^G and σ^K, respectively. Thus, a transcriptional hierarchy governs sporulation in *C.* difficile with downstream factors depending on the activation of upstream sigma factors. Since genes in the regulons of all four sporulation-specific sigma factors were under-expressed in Δ*camA* relative to wild type, whereas a relatively small subset of Spo0A regulon genes exhibited this pattern of regulation (Extended Data Fig. 9b, Supplementary Table 8e), loss of CamA likely affects early events during sporulation.

To identify the regulatory stage of sporulation that CamA-mediated DNA methylation specifically impacts, we used qRT-PCR to analyze the expression of genes encoding Spo0A, the sporulation-specific sigma factors[31,35], and genes in their individual regulons[31,35,36]. Consistent with our RNA-Seq analyses, Spo0A regulon genes, *spo0A, sigF,* and *sigE*[31,36], were expressed at similar levels between wild type and Δ*camA* at both 9 and 11 h, implying that the Δ*camA* mutant activates Spo0A at levels similar to wild type. In contrast, σ^F and σ^E regulon genes, *spoIIQ* and *spoIVA*[31,37], respectively, were under-expressed in Δ*camA* relative to wild type (Fig. 5c). Reduced SpoIIQ and SpoIVA levels were observed in Δ*camA* by western blot, confirming the transcriptional analyses (Extended Data Fig. 9c). Based on the hierarchical organization of the sporulation regulatory cascade, σ^F activation would appear to be the earliest sporulation stage affected by CamA. This conclusion is supported by our morphological analyses, since fewer Δ*camA* cells progress to engulfment (a process that requires both σ^F and σ^E activation[38]) than wild type (Fig. 2c), whereas similar numbers of Δ*camA* and wild-type 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 under-expressed in Δ*camA* cells could be dually regulated by Spo0A and σ^F. For example, the *spoIIR*[35] gene, which encodes a signaling protein required for σ^E activation, is activated by both Spo0A and σ^F [31,34].

In vivo impacts of the *camA* mutation

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

To test whether loss of CamA leads to defects in virulence, we compared Δ*camA* and wild type in a hamster model of infection. Clindamycin-treated golden Syrian hamsters are highly susceptible to the effects of the *C. difficile* toxins and thus represent a model of acute disease. Groups of 6 hamsters were inoculated by oral gavage with spores of the wild-type, Δ*camA*, and Δ*camA*-*C* strains. These *C. difficile* strains elicited diarrheal symptoms and weight loss in this model, and we observed no difference in animal survival times post inoculation (Fig. 6b). This result is consistent with the observation that the wild-type, 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 *C. difficile*'s ability to persist within the host intestinal tract.

Additional functional impacts of the *camA* mutation

Considering the high conservation of *camA* across *C. difficile* genomes, we asked if some additional phenotypes could be impacted by the gene's inactivation. While analyzing images of sporulating *C. difficile*, we noticed that Δ*camA* mutant cells appeared to be shorter on average than wild-type cells. To test this possibility, we measured the lengths of wild-type and Δ*camA* cells during broth culture and sporulation, and found that Δ*camA* cells were ~15% shorter than wild-type cells (Figs. 6c, d) even though no difference in growth was observed (Extended Data Fig. 5b). Interestingly, genes encoding putative cell wall remodeling enzymes, were over-expressed in the Δ*camA* mutant relative to wild type 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 (wild type vs. 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 to DE genes in conditions favoring the production of biofilm on a solid substrate[39] (Fig. 6e). Motivated by this significant overlap, we performed crystal violet staining assays of adherent biofilm biomass, and consistently observed that the Δ*camA* mutant produced more biofilm than wild type (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 DE during infection in different murine gut microbiome compositions[40] (Extended Data Fig. 10a, Supplementary Table 8f). Lastly, significant overlaps were found when comparing with DE genes obtained from murine gut isolates at increasing time points after infection[41] (Extended Data Fig. 10b, Supplementary Table 8f). Collectively, these integrative analyses provide additional 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 work to elucidate the mechanisms underlying the functional roles of CAAAAA methylation in *C. difficile* pathogenicity.

Discussion

C. difficile is responsible for one of the most common hospital-acquired infections and classified by the US Centers for Disease Control and Prevention as an urgent healthcare risk with significant morbidity and mortality[42]. Because *C. difficile* infection is spread by bacterial spores found within feces, extensive research has been devoted to better understand the genome of this critical 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*) conserved across all isolates (and in another ~300 published *C. difficile* genomes) sharing 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. Since enumeration of *C. difficile* recovered in feces of the infected animals reflects the number of *C. difficile* spores in the gut, the reduced burden of Δ*camA* in the mouse may 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[43]. The comparable virulence between Δ*camA* and wild type in the hamster model suggests that DNA methylation does not impact toxin-mediated disease. However, due to the pleiotropic nature of the MTase it remains possible that multiple factors contribute to the more pronounced effect observed in the mouse model.

The highly conserved *camA* and its flanking genes across *C. difficile* genomes suggest that additional phenotypes may be regulated by CamA beyond sporulation. Consistently, CAAAAA sites were overrepresented 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, is that its loss reduces cell length and results in statistically significant overlap between transcriptional signatures identified in our study (wild type vs Δ*camA* mutant) and those of others observed during the *in vivo* colonization and biofilm formation (Fig. 6e, Extended Data Figs. 10a, b).

The fact that *camA* is a solitary MTase gene without a cognate restriction gene further supports a view that widespread methylation in bacteria has a functional importance beyond that attributed to R-M systems. Previously, the most extensively characterized 6mA MTase was Dam targeting GATC in *E. coli*. Dam plays multiple important functions and is essential in some pathogens[12]. However, since it is conserved in the large diversity of Enterobacteria, it was not considered a promising drug target. In contrast, the uniqueness of *camA* in all *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 since gut dysbiosis potentiates *C. difficile* infection[44,45]. In addition, since this MTase seemingly does not impact the general fitness of *C. difficile*[23], a drug specifically targeting it may be developed with a lower chance for resistance.

Considering the large number of genes differentially expressed in the Δ*camA* mutant, the functional impact of CAAAAA methylation is likely 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[7,8,12,46], so the competition between transcription factors and MTases may form an epigenetic switch to turn on/off a gene.

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 analyze bacterial epigenomes is limited by a bottleneck in integratively characterizing methylation events, methylation motifs, transcriptomic data, and functional genomics data. In this regard, this work represents 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.

Materials and Methods

Clostridioides difficile isolates and culture

36 clonal *C. difficile* isolates from infected fecal samples were obtained using protocols developed in an ongoing Pathogen Surveillance Program at Mount Sinai Hospital (Supplementary Table 1). Additionally, 9 fully sequenced and assembled *C. difficile* genomes were retrieved from Genbank Refseq (ftp://ftp.ncbi.nih.gov/genomes, last accessed in 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 prior to analysis. All stool samples underwent culture for *C. difficile* using an ethanol shock culture method[47]. Briefly, 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 vortex mixed and incubated at room temperature for 20 min. A loopful was then cultured onto *C. difficile* selective agar (CDSA, Becton Dickinson, Franklin Lakes, NJ) and the plates were incubated anaerobically at 37 °C for up to 72 h. A single colony was subcultured onto a Trypticase™ soy agar with 5% defibrinated sheep blood plate (TSA II™, Becton Dickinson, Franklin Lakes, NJ) and incubated anaerobically at 37 °C for 48 h, after which colonies giving the characteristic *C. difficile* odor and fluorescence under UV illumination were obtained and confirmed by MALDI on a Brucker biotyper. For long-term storage, individual colonies were emulsified in tryptic soy broth containing 15% glycerol and stored at −80 °C.

Single-molecule real-time (SMRT) sequencing

Primer was annealed to size-selected (>8 kb) SMRTbells with the full-length libraries (80 ºC for 2 min and 30 s followed by decreasing the temperature by 0.1 ºC increments to 25 ºC). The polymerase-template complex was then bound to the P6 enzyme using a ratio of 10:1 polymerase to SMRTbell at 0.5 nM for 4 h at 30 ºC and then held at 4 ºC until ready for magbead loading, prior to sequencing. The magnetic bead-loading step was conducted at 4 ºC for 60 min per manufacturer's guidelines. The magbead-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 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 of a putative motif are examined by comparing the ipdR distribution of each variation with non-methylated motifs.

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

v2.8.0 and subsequently mapped to the *C. difficile* 630 reference genome using Bowtie2 v2.2.9[48] in paired-end mode. The resulting SAM files were converted to BAM format (with removal of unmapped reads and PCR duplicates), and sorted using SAMTOOLS v1.9[49]. To assess coverage, sequence depths were computed using the genomeCov function of BEDTOOLS v2.26.0[50] for each strand separately. Variant sites were called from the aligned reads using the *mpileup* and *bcftools* tools in SAMTOOLS.

Identification of defense systems

Identification of R-M systems was performed as previously described[51]. Briefly, curated reference protein sequences of Types I, II, IIC and III R-M systems and Type IV REases were downloaded from the data set 'gold standards' of REBASE[52] (last accessed in November 2016). All-against-all searches were performed for REase and MTase standard protein sequences retrieved from REBASE using BLASTP v2.5.0+ (default settings, *e* value $< 10^{-3}$). The resulting *e* values were log-transformed and used for clustering into protein families by Markov Clustering (MCL) v14–137[53]. Each protein family was aligned with MAFFT v7.305b[54] using the E-INS-i option, 1,000 cycles of iterative refinement, and offset 0. Alignments were visualized in SEAVIEW v4.6.1[55] 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 v3.0 suite[56] (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) v1.2[57] with default parameters. For CRISPR spacer homology search, we considered as positive hits those with at least 80% identity. For *cas* gene identification, we obtained Cas protein family HMMs from the TIGRFAM database[58] v15.0 and PFAM families annotated as Cas families (downloaded from ftp://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 (Abi) systems were searched with the PFAM profiles PF07751, PF08843, and PF14253 (last accessed in January 2018). Bacteriophage Exclusion (BREX) systems were searched using PFAM profiles for the core genes *pglZ* (PF08655) and *brxC*/*pglY* (PF10923), and specific PFAM profiles for each BREX type as indicated previously[59]. DISARM systems were identified using the PFAM signature domains (PF09369, PF00271, PF13091) belonging to the core gene triplet characteristic of this system[60]. To search for prokaryotic Argonaute (pAgo) genes we built a dedicated HMM profile based on a list of 90 Ago-PIWI proteins[61]. Searches for the ensemble of newly found antiphage systems were performed using the list of PFAM profiles published by the authors[62]. Type II toxin-antitoxin (T-A) systems were detected using the TAFinder tool[63] with default parameters. Matches of CRISPR spaces were performed against well-known *C. difficile* phages: five siphophages (ϕCD111 (NC_028905.1), ϕCD146 (NC_028958.1), ϕCD38–2 (NC_015568.1), ϕCD6356 (NC_015262.1), 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), ϕCDHM13 (NC_029116.1)), five medium-tail myophages (ϕMMP03 (NC_028959.1), ϕCDMH1 (NC_024144.1), ϕC2 (NC_009231.1), ϕCD119 (NC_007917.1), ϕCDHM19 (NC_028996.1)), and four long-tail myophages

(ϕCD27 (NC_011398.1), ϕMMP02 (NC_019421.1), ϕCD505 (NC_028764.1), ϕMMP01 (NC_028883.1)).

Identification and classification of prophages, conjugative/mobilizable elements and integrons

Prophages were detected using Phage Finder v2.1[64] under strict mode, and PHASTER[65] under default settings. We took the common hits obtained by both programs, as well as those very few cases (~10% of the hit list) corresponding to complete prophages predicted by just one of the programs. All elements smaller than 18 kb, or lacking matches to core phage proteins (e.g. terminase, capsid, head, tail proteins) were removed. Integrons were searched with IntegronFinder[66] under default settings. The identification of genes encoding the functions related to conjugation in integrative conjugative elements (ICEs) was performed as previously described[67]. Briefly, an element was considered as 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). In the case of integrative mobilizable elements (IMEs), they 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[68] v3.8.31 (default parameters). Since at this evolutionary distance the DNA sequences provide more phylogenetic signal than protein sequences, we back-translated the alignments to DNA. Poorly aligned regions were removed with BMGE[69] v1.12. The tree was computed with RAxML[70] v8.00 under the GTR model and a gamma correction (GAMMA) for variable evolutionary rates. 100 bootstraps were performed 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 methodology previously published[71]. Briefly, a preliminary list of orthologs 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 <80% similarity in amino-acid sequence or >20% difference in protein length were discarded. This list of orthologs was then refined for every pairwise comparison using information on the conservation of gene neighborhood. Positional orthologs were defined as bi-directional best hits adjacent to at least four other pairs of bi-directional best hits within a neighborhood of 10 genes (five upstream and five downstream). The core-genome of each clade was defined as the intersection of pairwise lists of positional orthologs. 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 with BLASTP (default parameters). We then used the *e*-values ($<10^{-4}$) of the BLASTP output to cluster them using SILIX[72] v1.2.11. 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 >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[73] by the Heap's power law $n = k \cdot N^{-\alpha}$, 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 closed, i.e. sampling more genomes will not affect its size. For $\alpha < 1$ ($0 < \delta < 1$) the pan-genome remains open and 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 LCBs obtained by progressiveMauve were used) using ClonalFrameML[74] v10.7.5 and Geneconv[75] v1.81a. The first used a predefined tree (*i.e.* the specie's 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[74]. Mean patristic branch lengths were computed with the R package "ape"[76] v3.3, and transition/transversion ratios were computed with the R package "PopGenome"[77] v2.1.6. 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 sites with missing data.

Reconstruction of the evolution of gene repertoires

We assessed the dynamics of gene family repertoires using Count[78] (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 with default parameters, assuming a Poisson distribution for the family size at the tree root and uniform duplication rates. One hundred rounds of rate optimization were computed 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 HGT 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 HGT 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*ermΔpyrE* parental strain was used for *pyrE*-based allelic-coupled exchange (ACE[79]). See Supplementary Table 5a for a list of *C. difficile* and *E. coli* strains. *C. difficile* strains were grown from frozen stocks on brain heart infusion media (BHIS)[80] plates supplemented with taurocholate (TA, 0.1% w/v; 1.9 mM), kanamycin (50 μg/mL), and cefoxitin (8 μg/mL) as needed. For ACE, *C. difficile* defined media (CDDM)[81] was supplemented with 5-fluoroorotic acid (5-FOA) at 2 mg/mL and uracil at 5 μg/mL. Cultures were grown at 37 °C under anaerobic conditions using 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, shaking at 225 rpm in Luria-Bertani broth (LB). The media was supplemented with chloramphenicol (20 μg/mL) and ampicillin (50 μg/mL) as needed.

E. coli strain construction

Primers used in this manuscript are listed 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[82]. 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 fashion 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*ermΔpyrEΔcamA* using uracil and 5-fluoroorotic acid to select for plasmid excision as previously described[83]. The flanking primer pair #2274 / 2279 was used to screen for the *camA* deletion as shown in Extended Data Fig. 5a (primers are provided in Supplementary Table 5b). Colonies that appeared to harbor 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 by using CDDM plates to select for restoration of the *pyrE* locus via recombination[83]. 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 100x Zeiss Plan Neofluar objective (1.3 NA) at each

timepoint. Cell length was calculated using the MicrobeJ plugin for Fiji/ImageJ[84]. Image thresholding was done 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-max, Width: 0.5–1 μm) and contours were generated using an interpolated rod-shaped method. Cell length data was exported from MicrobeJ and analyzed using Prism 8 (Graph-pad).

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} between 0.35 and 0.75, and then 120 μL of this culture was spread onto 70:30 plates (40 mL). Sporulating cultures were harvested into phosphate-buffered saline (PBS), the sample was pelleted, and sporulation levels were visualized by phase-contrast microscopy as previously described[85].

Fluorescence microscopy

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

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]. Briefly, 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, 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 optical density at 600 nm was measured. Spore purification yields were determined from three independent spore preparations. Statistical significance was determined using a one-way ANOVA and Tukey's test.

Heat resistance assay

Heat-resistant spore formation was measured in sporulating *C. difficile* cultures after 20–24 h as previously described[85]. The heat resistance (H_{RES}) efficiency represents the average ratio of heat-resistant colony forming units (CFUs) to total CFUs for a given strain relative to the average ratio determined for wild type. H_{RES} was determined based on the average H_{RES} values for a given strain in three biological replicates. Statistical significance was determined using a one-way ANOVA and Tukey's test.

Germination assay

Germination assays were performed as previously described[33]. Spores (0.35 OD_{600} units, corresponding to ~1×10^7) were resuspended in 100 μl of water, and 10 μL of this mixture was removed for 10-fold 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 wild-type spores for three biological replicates. Statistical significance was determined by performing a one-way ANOVA on natural log-transformed data using 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 (0.75 OD_{600} units, corresponding to ~2×10^7 spores) were re-suspended in 190 μL water. 90 μL were then 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] v3.1.0. 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 under 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 with a previous developed framework[27] involving a multi-scale representation (MSR) of genomic signals. We created a binary genomic signal for motif content, which was 1 at motif positions, and 0 otherwise. 50 length scales were used. Pruning parameter values were set to default and the *P* value threshold 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 by the progressiveMauve program[88] v2.4.0 with default parameters. Since 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 locally collinear blocks (LCBs) of size ≥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. The former correspond to those situations where the CAAAAA motif was absent in at least one genome. Typically, these correspond to regions containing MGEs or unaligned repetitive regions.

Identification of transcription factor binding sites, and transcription start sites

Identification of transcription factor binding sites (TFBS) was performed by retrieving *C. difficile* 630 regulatory sites in FASTA format from the RegPrecise database[91] (last accessed July 2017). These were converted to PSSMs using in-house developed scripts. This led to a total of 21 PSSMs pertaining to 14 distinct transcription factor families (Supplementary Table 7b). Matches between these matrices and *C. difficile* genomes was performed with MAST[92] (default settings). MAST output was filtered on the basis of *P* value. Hits with *P* value $<10^{-9}$ were considered positive, while hits $>10^{-5}$ were considered negative. Hits with intermediate *P* values were only considered 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 (which have shorter binding sites), we considered positive those hits with *P* values $<10^{-8}$. Transcription start sites (TSSs) were predicted with Parseq[93] under 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 kept 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 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 wild type and *ΔcamA* growing on 70:30 sporulation media after 9 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 work[31]. For analyses of the mid-log and early stationary phase cultures, overnight cultures of wild type 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 harvested 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 last treatment being on column using the Qiagen RNeasy kit. Samples were tested for genomic DNA contamination using quantitative PCR 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 harvested from sporulating cultures using the Ambion MICROB*Express* 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 qRT-PCR analyses were harvested in triplicate from a separate set of three biological replicates that were grown identically to the cultures used for RNA-Seq analyses. The RNA was processed similarly except that mRNA

enrichment was done using a MICROB*Express*, and the DNase-treated RNA samples for qRT-PCR analyses were tested for genomic DNA contamination using quantitative PCR for *rpoB*.

RNA sequencing, read alignment, and differential expression analysis

Purified RNA was extracted from three biological replicates of sporulating (9, 10.5 h) and exponential and stationary grown 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[31]. RNA sequencing was performed on a HiSeq 2500, yielding an average of 29.4 (±4.5, sd) million 100-bp single-end reads per sample (exponential and stationary growth timepoints) and 26.9 (±4.3, sd) million 150-bp paired-end reads per sample (sporulation time points). Read quality was checked using FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). We used Trimmomatic[94] v0.39 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 data set using SortMeRNA[95] v2.1, based on the SILVA 16s and 23s rRNA databases[96], and Rfam 5s rRNA database[97]. The resulting non-rRNA reads were mapped to the *C. difficile* 630 reference genome using BWA-MEM v0.7.17-r1198[98]. The resulting bam files were sorted and indexed with SAMTOOLS, and read assignment was performed with featureCounts[99] v1.6.4 (excluding multi-mapping and multi-overlapping reads). A gene was included for differential expression analysis if it had more than one count in all samples. Normalization and differential expression testing were performed using the Bioconductor package DESeq2 v1.18.1[100]. Genes with a false discovery rate (FDR) < 0.05 and $|\log_2 FC| > 1$ were called as differentially expressed. Functional classification of genes was performed using the DAVID online database (https://david.ncifcrf.gov)[101]. GO annotation terms with a gene count ≥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[102] v5.2 and Panther[103] v14. Briefly, for Blast2GO, we ran BLASTX searches of the *C. difficile* 630 genome against the entire GenBank bacterial protein database (as of 09/2018). 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 in the website (http://www.pantherdb.org). To assess the significance of the intersection between multiple datasets of differentially expressed genes (typically observed during *C. difficile* colonization and infection), we collected gene-expression data from *in vivo* and *in vitro* studies[39–41], in which key factors for gut colonization (*e.g.*, time post-infection, antibiotic exposure, and spatial structure (planktonic, biofilm growth)) were tested. Differentially expressed genes were called under the same conditions as described above. Statistical analyses and graphical representation of multi-set intersections was performed with the R package *SuperExactTest*[104].

Quantitative real-time PCR (qRT-PCR)

Transcript levels were determined from cDNA templates prepared from the three biological replicates described above. Gene-specific primer pairs are provided in Supplementary Table 5b. qRT-PCR was performed as described[32], 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 harvested and processed for immunoblotting as previously[86]. Total protein in each sample was quantified using the Pierce 660nm protein assay with the ionic detergent compatibility reagent (Thermo Fisher) and 5 μg of protein was loaded for each sample. σ^F, σ^E, and Spo0A were resolved on 15% SDS-PAGE gels while SpoIIQ and SpoIVA were resolved on 12% SDS-PAGE gels. Protein was transferred to PVDF membranes, which were subsequently probed with rabbit (σ^F, σ^E, SpoIIQ) and mouse (Spo0A, SpoIVA) polyclonal primary antibodies and α-rabbit IR800/α-mouse IR680 secondary antibodies (LI-COR). Blots were imaged on the LiCor Odyssey CLx. 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 by centrifugation, suspended in SDS-PAGE buffer, and boiled for 10 min. 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 a mouse α-TcdA primary antibody (Novus Biologicals) and goat anti-mouse IgG conjugated with IR800 (Thermo Fisher)[105].

Animal infection studies—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. Animals considered 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.

Murine 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[106,107]. 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), gentamicin (35 μg/ml), colistin (850 units/ml), vancomycin (45 μg/ml), and metronidazole (215 μg/ml)) in their water *ad libitum* seven days prior to inoculation for three days, followed by a single intra-peritoneal dose of clindamycin (10 mg/kg body weight) 2 days prior to inoculation. Mice were randomly assigned into

groups, with two mice assigned to the mock condition and six mice (3 male, 3 female) to each infection condition. The experiment was independently repeated to assess consistency of the data. The data from the experiments were combined for analysis for a total of 12 mice (6 male, 6 female) in each infection condition. Mice were inoculated with 10^5 spores by oral gavage. Mock-inoculated animals were included as controls. Cage changes were performed every 48 h post-inoculation. Fecal samples were collected every 24 h for seven days post-inoculation. Dilutions were plated on BHIS-agar containing 0.1% of the germinant taurocholate to enumerate spores as colony forming units (CFU) per gram of feces.

Hamster model—The above strains were tested in Syrian golden hamster strain LVG (*Mesocricetus auratus*; Charles River Laboratories) as described previously[108]. Hamsters were randomly assigned into groups, with two assigned to the mock condition and six (3 male, 3 female) to each infection condition. Hamsters were administered a single dose of clindamycin (30 mg/kg body weight) by oral gavage, then inoculated with approximately 5,000 spores of the above strains 5 days later. Hamsters were monitored for weight loss and diarrheal symptoms and were considered moribund after 15–20% weight loss from maximum body weight, with or without concurrent diarrhea.

Biofilm assays—Biofilm assays were done as previously described[109]. Briefly, overnight cultures of *C. difficile* were diluted 1:100 in BHIS-1% glucose-50 mM sodium phosphate buffer (pH 7.5) in 24-well polystyrene plates. After 24 hours of growth at 37 °C, supernatants were removed, the biofilms were washed once with PBS and then stained for 30 minutes with 0.1% (w/v) crystal violet. After 30 minutes, 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, with each strain assayed in quadruplicate in each experiment.

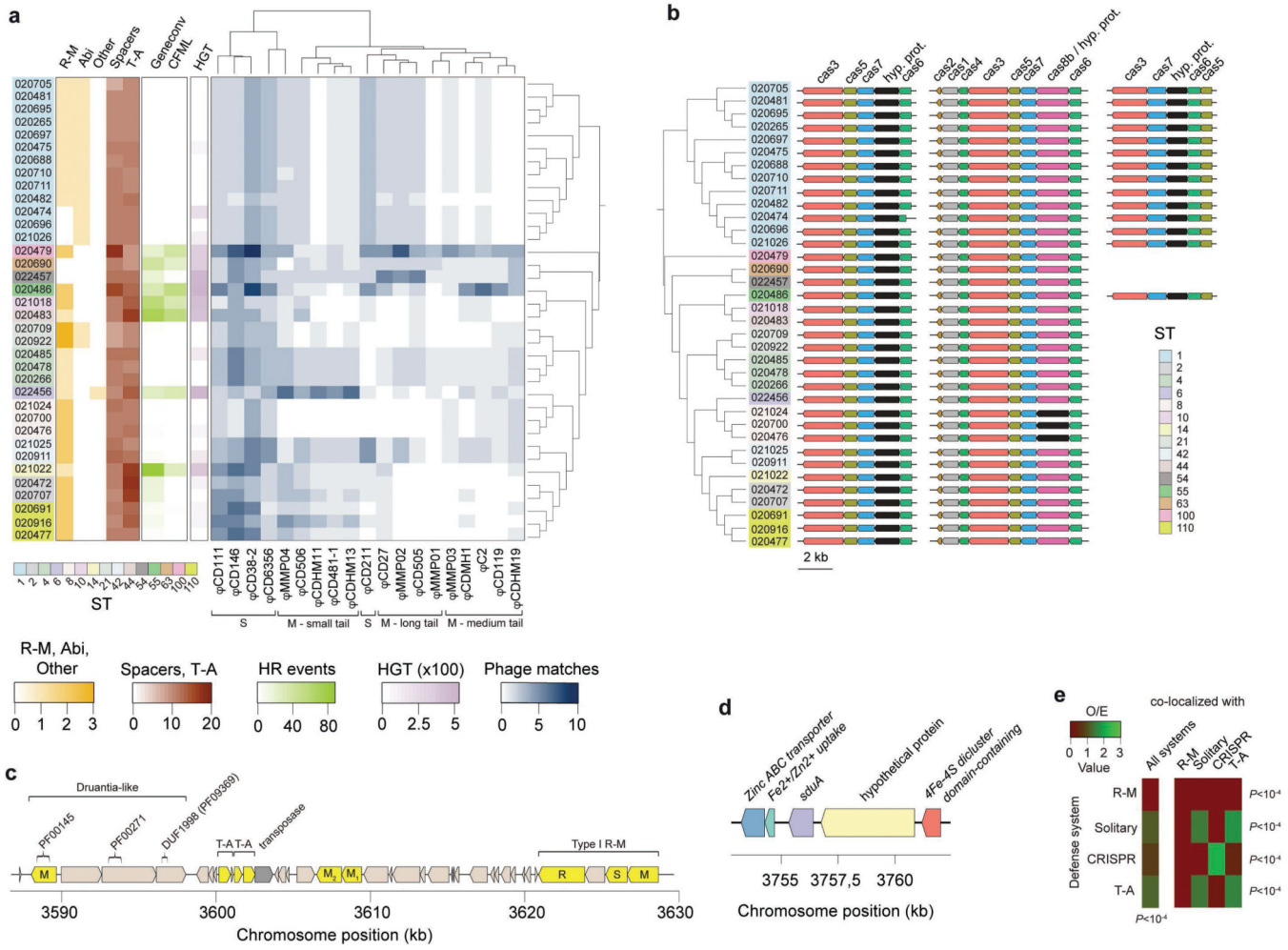
Data availability

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

Code availability

Scripts and a tutorial supporting all key analyses of this work are publicly available as a package named **B**acterial **E**pigenome **A**nalysis **S**ui**T**e (BEAST) at http://github.com/fanglab/.

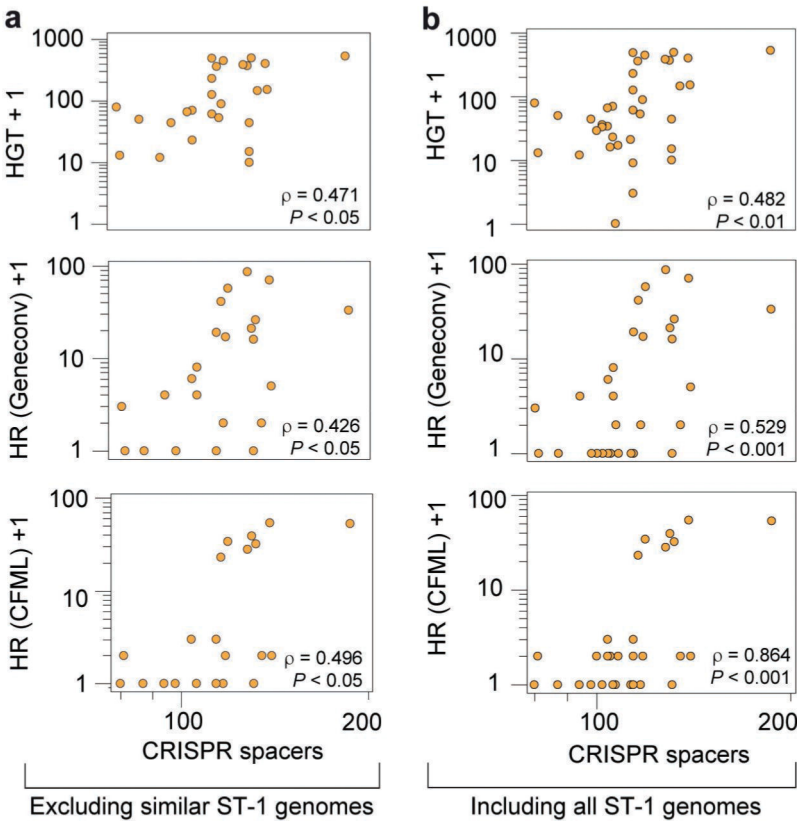
Extended Data



Extended Data Fig. 1. Multiple defense systems and gene flux control in *C. difficile*.

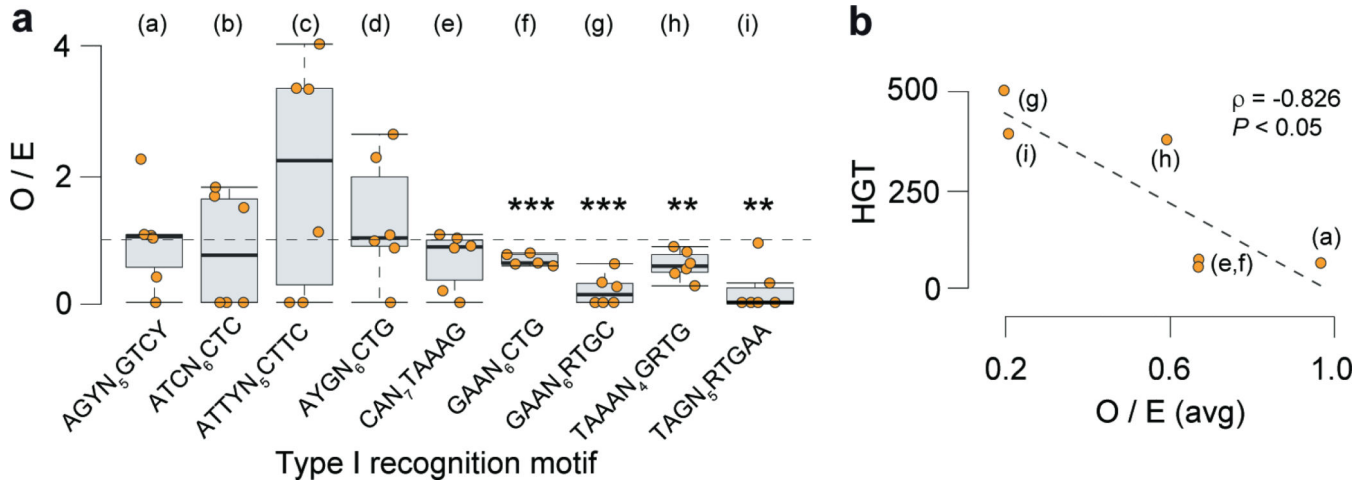
Multiple defense systems and gene flux control in *C. difficile*. (a) Heatmap aggregate depicts: abundance of defense systems (R-M, abortive infection (Abi), average number of spacers per CRISPR, toxin-antitoxin (T-A), and Shedu systems (other)), homologous recombination (HR) events (given by Geneconv 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[62] 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.



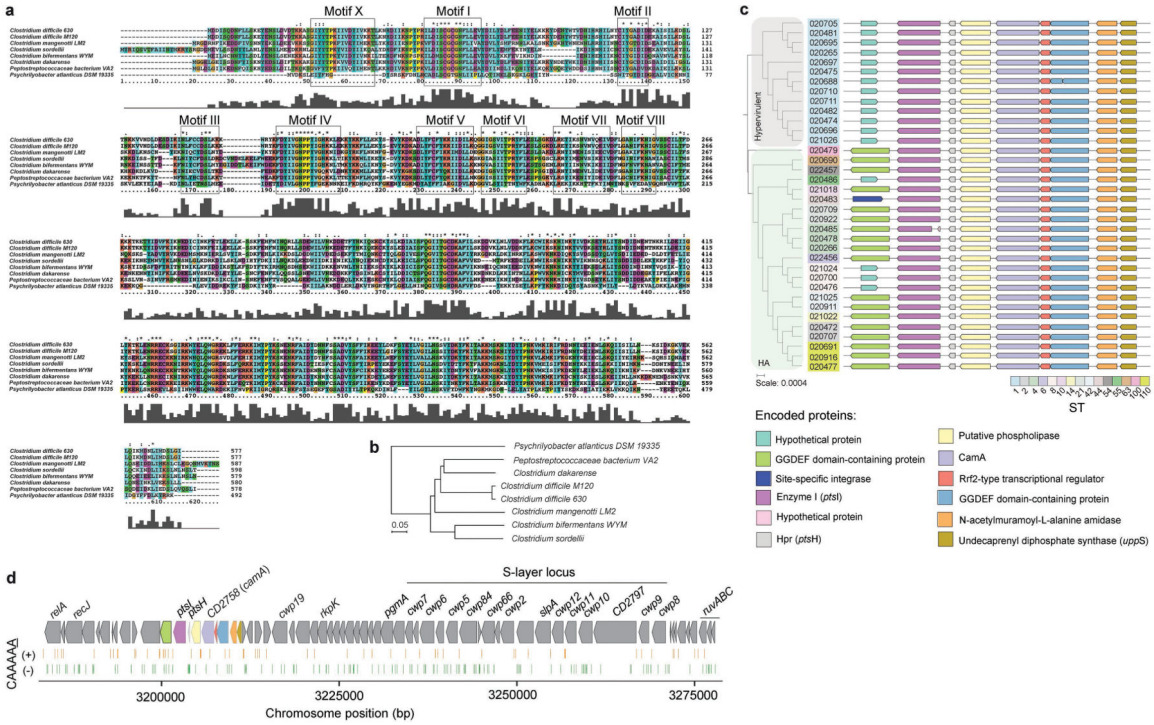
Extended Data Fig. 2. Relation between gene flux and CRISPR spacer content.

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, i.e., 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.



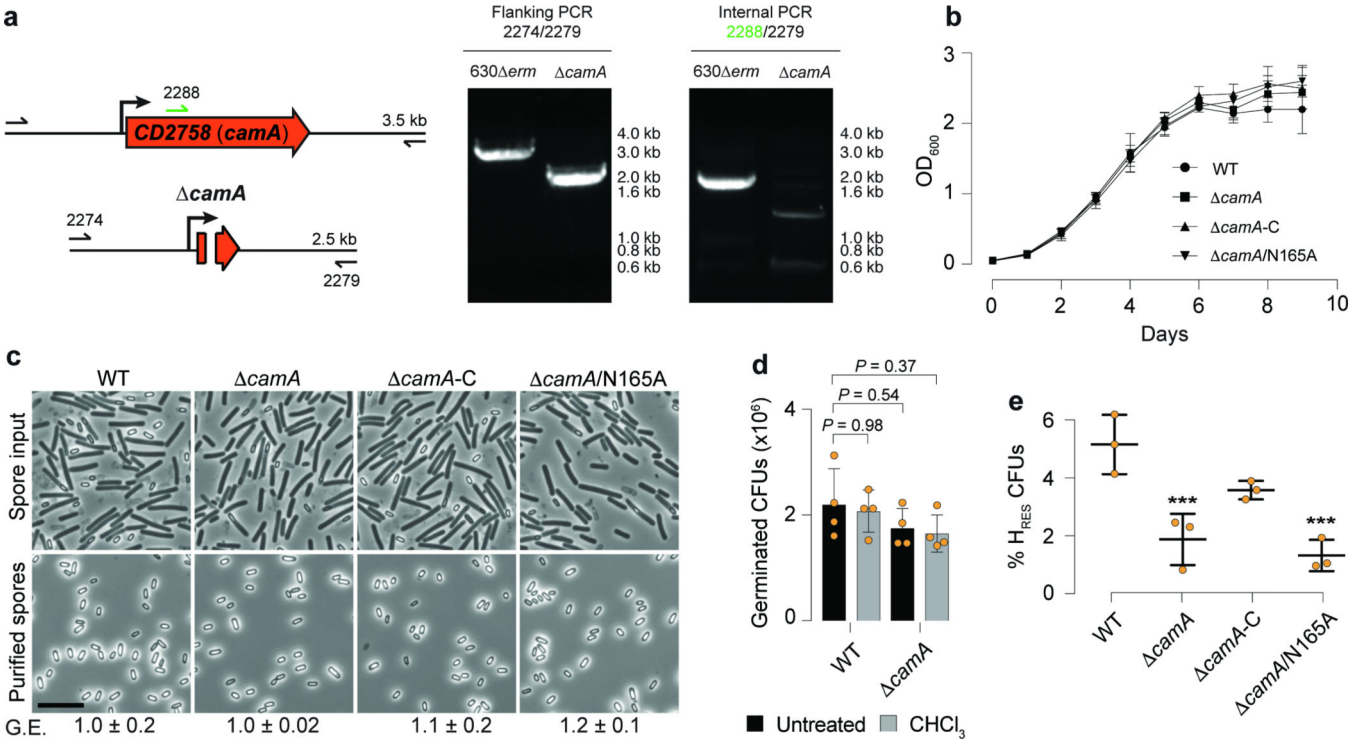
Extended Data Fig. 3. Interplay between Type I R-M systems and gene flux in *C. difficile*.

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 (i.e., 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.



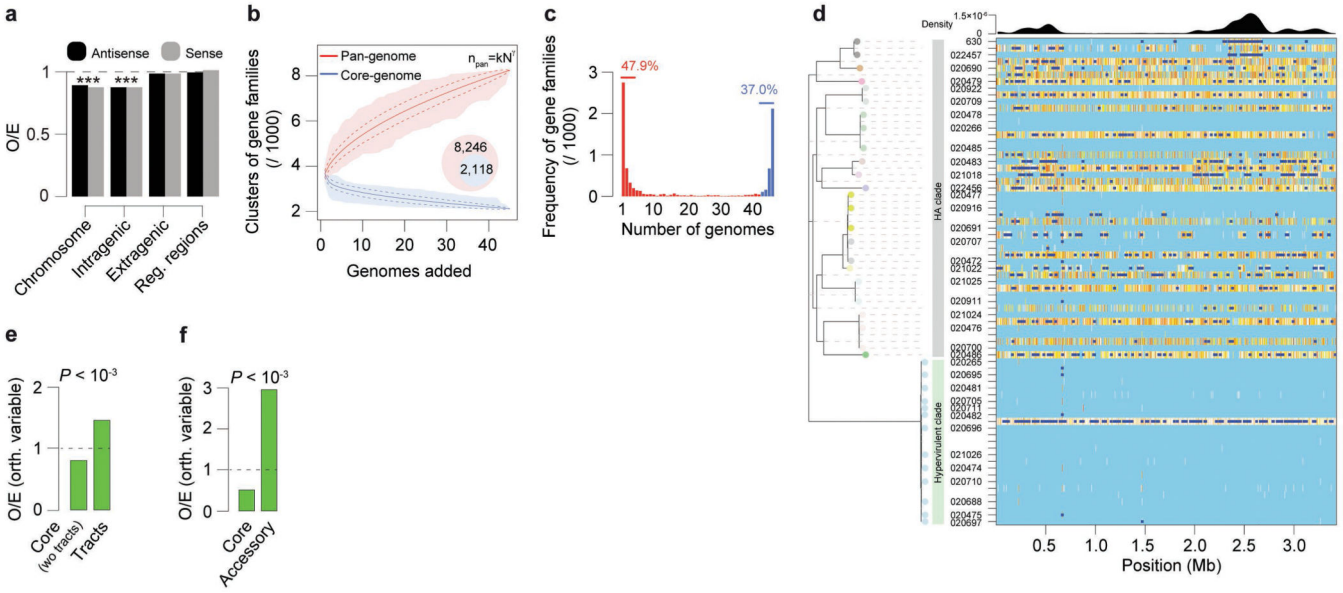
Extended Data Fig. 4. Genomic context and conservation of *camA*.

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.

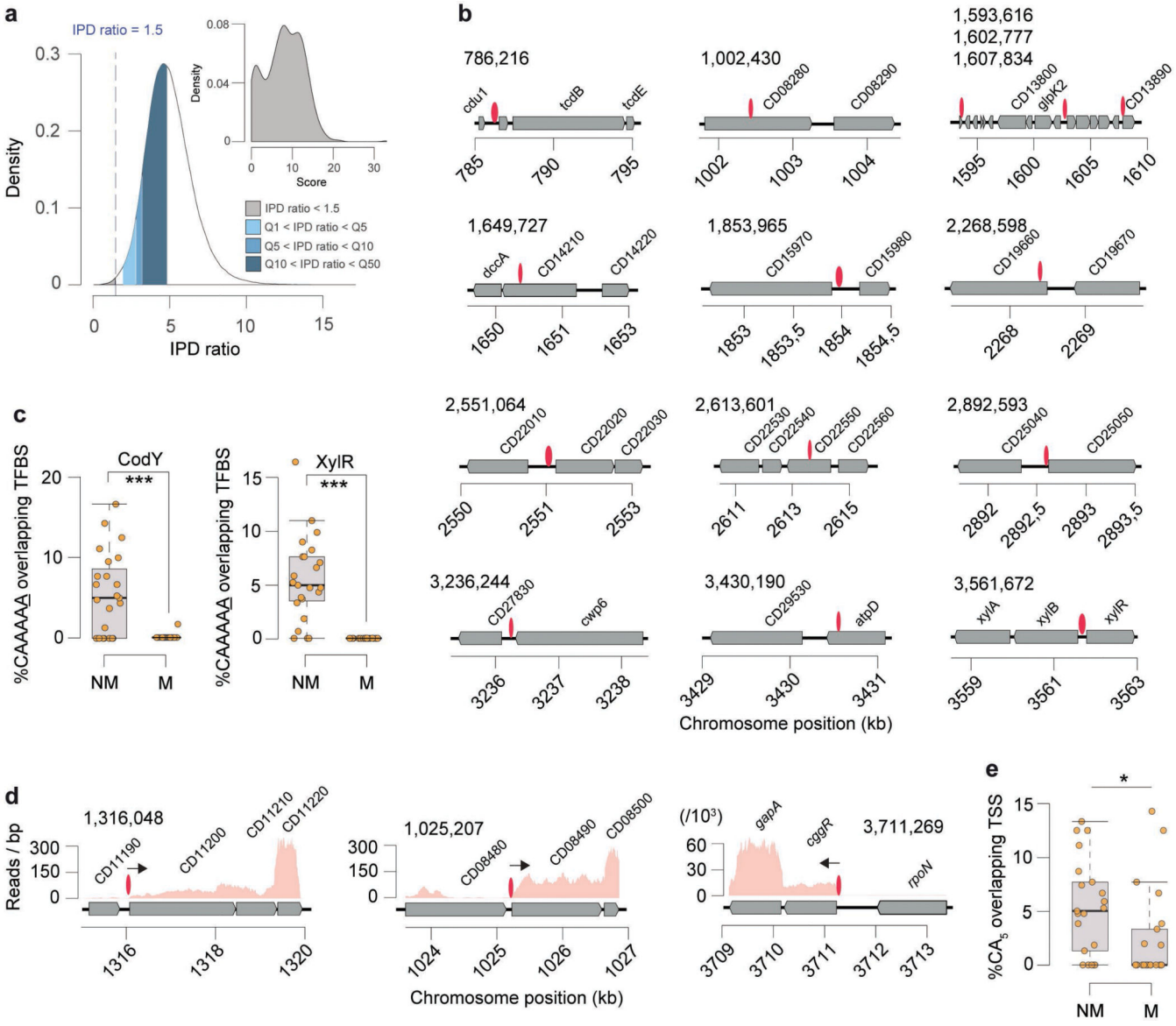
Δ*camA* construction, purified spore analyses, broth culture growth, and sporulation kinetics. (a) PCR to distinguish between wild-type *camA* and Δ*camA* using flanking primers and primers internal to the deletion. PCRs were performed twice independently. (b) Growth curves comparing wild-type *camA,* Δ*camA*, *ΔcamA-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.

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

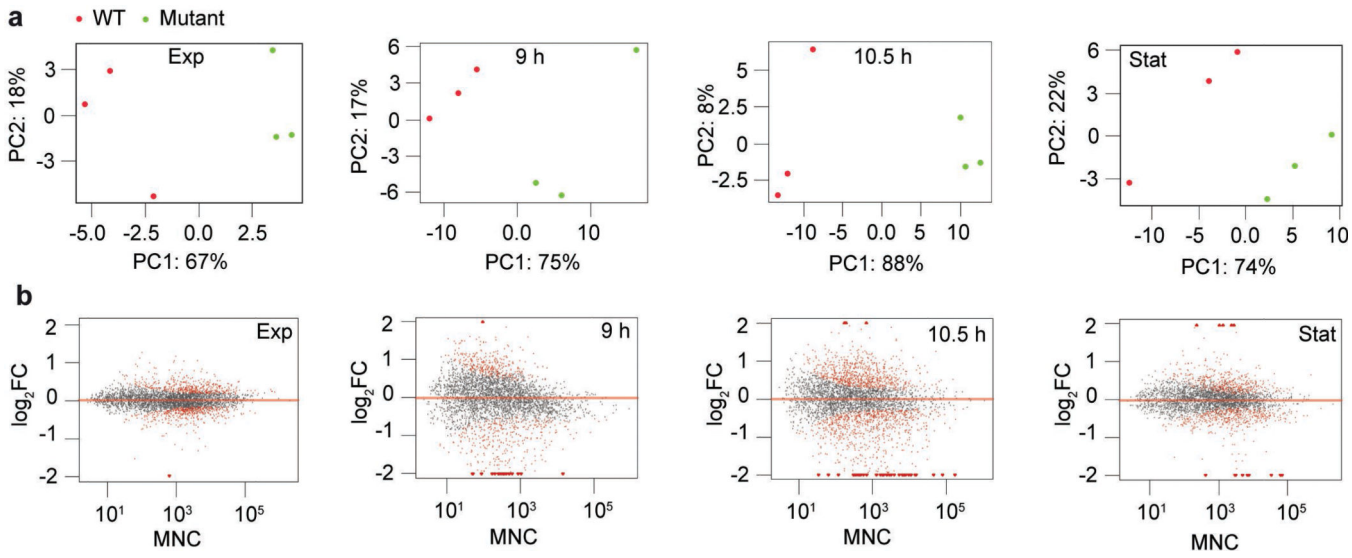
recombination tracts) (n = 770) and recombination tracts (n = 325), or (f) core (n = 1,095) and accessory genome (n = 1,415). P values correspond to the Chi-square test.



Extended Data Fig. 7. Non-methylated CAAAAA motif sites overlapping TFBS and TSS.

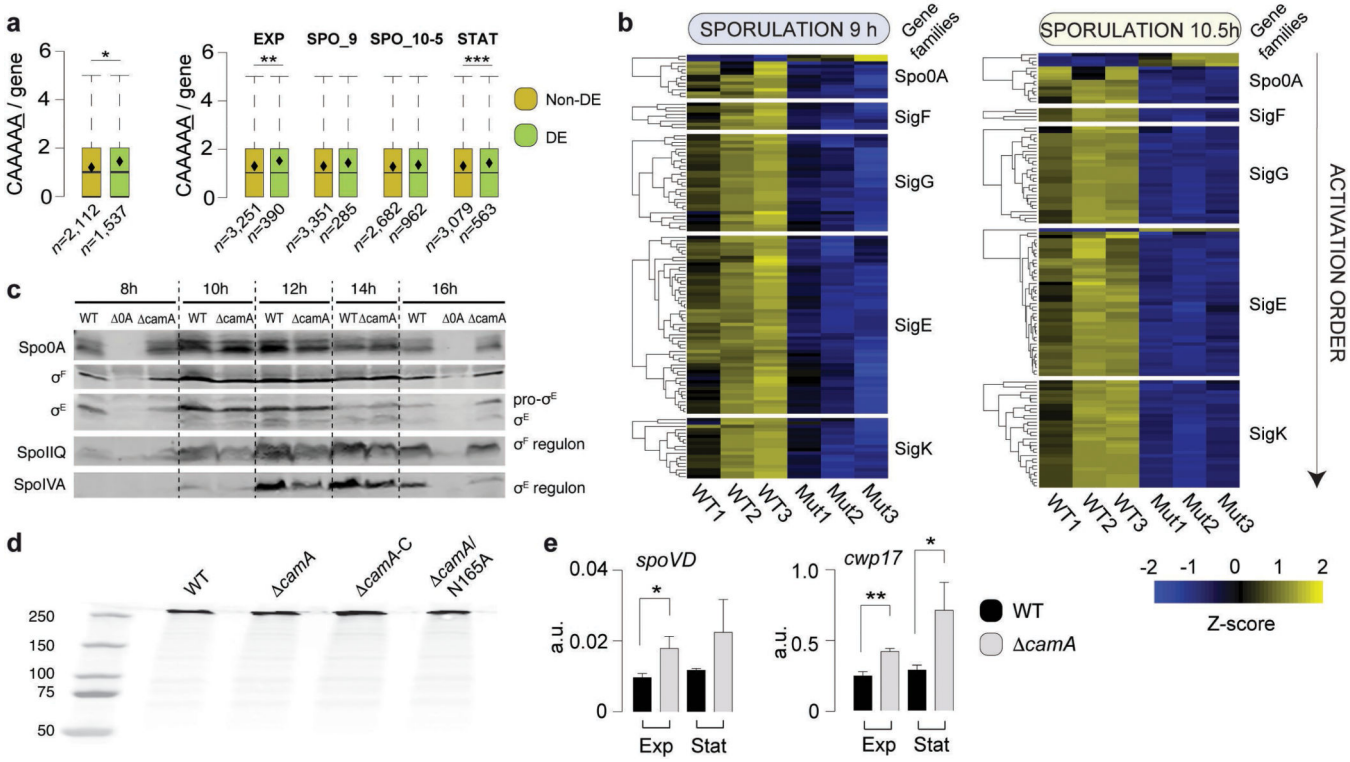
Non-methylated CAAAAA motif sites overlapping TFBS and TSS. (a) Interpulse duration ratio (ipdR) density distribution of the terminal adenine of CAAAAA. Motifs were considered as non-methylated if the terminal adenine had IPD ratios <1.5 (stippled line), coverage >20×, and methylation scores <20 (gray 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 25^{th} and 75^{th} 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).



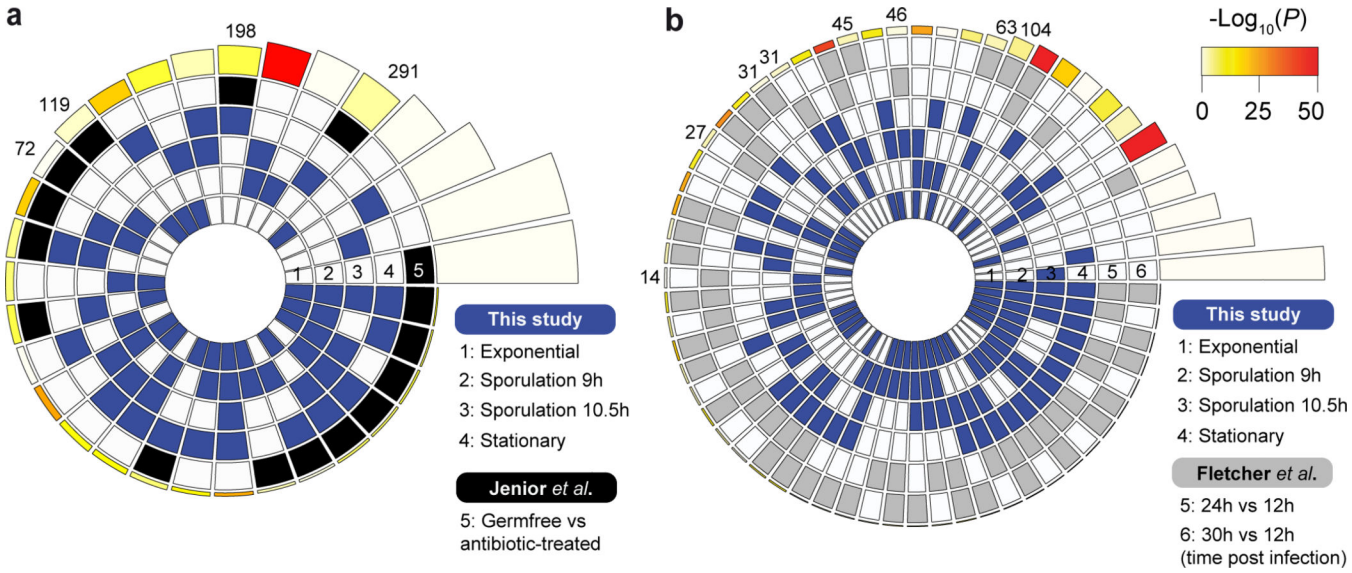
Extended Data Fig. 8. Principal Component Analysis (PCA) and MA-plots for RNA-seq data.

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.

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) qRT-PCR 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

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 murine gut microbiome compositions ($P < 10^{-6}$, one-tailed hypergeometric test implemented in *SuperExactTest*, Bonferroni adjusted), or (b) DE genes obtained from murine gut isolates at increasing time points after infection ($P < 10^{-4}$, one-tailed hypergeometric test implemented in *SuperExactTest*, Bonferroni adjusted).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We acknowledge Dr. Richard J. Roberts (New England Biolabs, Inc. USA) for his help with the prediction of R-M systems and orphan MTases in *C. difficile* genomes using REBASE Tools and for providing helpful 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 acknowledge Dr. Eduardo P.C. Rocha (Institut Pasteur, Paris, France) for critical reading and for providing helpful comments. The work was primarily funded by R01 GM114472 (G.F.) from the National Institutes of Health and Icahn Institute for Genomics and Multiscale Biology. In addition, the work was funded by NIH grants R01 AI119145 (H.v.B and A.B.), R01 AI22232 (A.S.) and R01 AI107029 (R.T.) a Hirschl Research Scholar award from the Irma T. Hirschl/Monique Weill-Caulier Trust (G.F.), a Pew Scholar in the Biomedical Sciences grant from the Pew Charitable (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. R.J.R.'s participation in this project was funded by New England Biolabs. This work 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.

References

1. Smits WK, Lyras D, Lacy DB, Wilcox MH, Kuijper EJ. Clostridium difficile infection. Nat. Rev. Dis. Primers 2, 16020 (2016). [PubMed: 27158839]
2. Sebaihia M, et al. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. Nat. Genet 38, 779–786 (2006). [PubMed: 16804543]
3. He M, et al. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. Nat. Genet 45, 109–113 (2013). [PubMed: 23222960]
4. Herbert M, O'Keeffe TA, Purdy D, Elmore M, Minton NP. Gene transfer into *Clostridium difficile* CD630 and characterisation of its methylase genes. FEMS Microbiol. Lett 229, 103–110 (2003). [PubMed: 14659549]
5. van Eijk E, et al. Complete genome sequence of the *Clostridium difficile* laboratory strain 630Deltaerm reveals differences from strain 630, including translocation of the mobile element CTn5. BMC Genomics 16, 31 (2015). [PubMed: 25636331]
6. Hargreaves KR, Thanki AM, Jose BR, Oggioni MR, Clokie MR. Use of single molecule sequencing for comparative genomics of an environmental and a clinical isolate of *Clostridium difficile* ribotype 078. BMC Genomics 17, 1020 (2016). [PubMed: 27964731]
7. Casadesus J, Low D. Epigenetic gene regulation in the bacterial world. Microbiol. Mol. Biol. Rev 70, 830–856 (2006). [PubMed: 16959970]
8. Low DA, Weyand NJ, Mahan MJ. Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. Infect. Immun 69, 7197–7204 (2001). [PubMed: 11705888]
9. Cohen NR, et al. A role for the bacterial GATC methylome in antibiotic stress survival. Nat. Genet. 48, 581–586 (2016). [PubMed: 26998690]
10. Manso AS, et al. A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. Nat. Commun 5, 5055 (2014). [PubMed: 25268848]
11. Atack JM, et al. A biphasic epigenetic switch controls immunoevasion, virulence and niche adaptation in non-typeable Haemophilus influenzae. Nat. Commun 6, 7828 (2015). [PubMed: 26215614]
12. Wion D, Casadesus J. N6-methyl-adenine: an epigenetic signal for DNA-protein interactions. Nat. Rev. Microbiol 4, 183–192 (2006). [PubMed: 16489347]
13. Oliveira PH, Touchon M, Rocha EP. Regulation of genetic flux between bacteria by restriction-modification systems. Proc. Natl. Acad. Sci. USA 113, 5658–5663 (2016). [PubMed: 27140615]
14. Flusberg BA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat. Methods 7, 461–465 (2010). [PubMed: 20453866]
15. Beaulaurier J, Schadt EE, Fang G. Deciphering bacterial epigenomes using modern sequencing technologies. Nat. Rev. Genet 20, 157–172 (2019). [PubMed: 30546107]
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). [PubMed: 23138224]
17. Murray IA, et al. The methylomes of six bacteria. Nucleic Acids Res. 40, 11450–11462 (2012). [PubMed: 23034806]
18. Davis BM, Chao MC, Waldor MK. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. Curr. Opin. Microbiol 16, 192–198 (2013). [PubMed: 23434113]
19. Smits WK. Hype or hypervirulence: a reflection on problematic C. difficile strains. Virulence 4, 592–596 (2013). [PubMed: 24060961]
20. Roberts RJ, et al. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nucleic Acids Res. 31, 1805–1812 (2003). [PubMed: 12654995]
21. Wust J, Sullivan NM, Hardegger U, Wilkins TD. Investigation of an outbreak of antibiotic-associated colitis by various typing methods. J. Clin. Microbiol 16, 1096–1101 (1982). [PubMed: 7161375]
22. Barra-Carrasco J, Paredes-Sabja D. Clostridium difficile spores: a major threat to the hospital environment. Future Microbiol 9, 475–486 (2014). [PubMed: 24810347]

23. Dembek M, et al. High-throughput analysis of gene essentiality and sporulation in *Clostridium difficile*. MBio 6, e02383 (2015). [PubMed: 25714712]
24. Donnelly ML, Fimlaid KA, Shen A. Characterization of *Clostridium difficile* spores lacking either SpoVAC or dipicolinic acid synthetase. J. Bacteriol 198, 1694–1707 (2016). [PubMed: 27044622]
25. Shen A, Fimlaid KA, Pishdadian K. Inducing and quantifying *Clostridium difficile* spore formation. Methods Mol. Biol. 1476, 129–142 (2016). [PubMed: 27507338]
26. Schbath S, Hoebeke M. R'MES: a tool to find motifs with a significantly unexpected frequency in biological sequences In: Advances in genomic sequence analysis and pattern discovery (ed^(eds Elnitsk i OP L, and Welch L). World Scientific (2011).
27. Knijnenburg TA, et al. Multiscale representation of genomic signals. Nat. Methods 11, 689–694 (2014). [PubMed: 24727652]
28. Lim HN, van Oudenaarden A. A multistep epigenetic switch enables the stable inheritance of DNA methylation states. Nat. Genet 39, 269–275 (2007). [PubMed: 17220888]
29. Ardissone S, et al. Cell cycle constraints and environmental control of local DNA hypomethylation in alpha-Proteobacteria. PLoS Genet. 12, e1006499 (2016). [PubMed: 27997543]
30. Cota I, Bunk B, Sproer C, Overmann J, Konig C, Casadesus J. OxyR-dependent formation of DNA methylation patterns in OpvABOFF and OpvABON cell lineages of *Salmonella enterica*. Nucleic Acids Res. 44, 3595–3609 (2016). [PubMed: 26687718]
31. Fimlaid KA, et al. Global analysis of the sporulation pathway of Clostridium difficile. PLoS Genet. 9, e1003660 (2013). [PubMed: 23950727]
32. Pishdadian K, Fimlaid KA, Shen A. SpoIIID-mediated regulation of sigmaK function during *Clostridium difficile* sporulation. Mol. Microbiol. 95, 189–208 (2015). [PubMed: 25393584]
33. Fimlaid KA, Shen A. Diverse mechanisms regulate sporulation sigma factor activity in the *Firmicutes*. Curr. Opin. Microbiol 24, 88–95 (2015). [PubMed: 25646759]
34. Saujet L, Pereira FC, Henriques AO, Martin-Verstraete I. The regulatory network controlling spore formation in *Clostridium difficile*. FEMS Microbiol. Lett 358, 1–10 (2014). [PubMed: 25048412]
35. Saujet L, et al. Genome-wide analysis of cell type-specific gene transcription during spore formation in *Clostridium difficile*. PLoS Genet. 9, e1003756 (2013). [PubMed: 24098137]
36. Rosenbusch KE, Bakker D, Kuijper EJ, Smits WK. *C. difficile* 630Deltaerm Spo0A regulates sporulation, but does not contribute to toxin production, by direct high-affinity binding to target DNA. PLoS ONE 7, e48608 (2012). [PubMed: 23119071]
37. Fimlaid KA, Jensen O, Donnelly ML, Siegrist MS, Shen A. Regulation of *Clostridium difficile* spore formation by the SpoIIQ and SpoIIIA proteins. PLoS Genet. 11, e1005562 (2015). [PubMed: 26465937]
38. Ribis JW, Fimlaid KA, Shen A. Differential requirements for conserved peptidoglycan remodeling enzymes during *Clostridioides difficile* spore formation. Mol. Microbiol. 110, 370–389 (2018). [PubMed: 30066347]
39. Maldarelli GA, et al. Type IV pili promote early biofilm formation by Clostridium difficile. Pathog. Dis 74, (2016).
40. Jenior ML, Leslie JL, Young VB, Schloss PD. *Clostridium difficile* colonizes alternative nutrient niches during infection across distinct murine gut microbiomes. mSystems 2, (2017).
41. Fletcher JR, Erwin S, Lanzas C, Theriot CM. Shifts in the gut metabolome and *Clostridium difficile* transcriptome throughout colonization and infection in a mouse model. mSphere 3, (2018).
42. Lessa FC, et al. Burden of *Clostridium difficile* infection in the United States. N. Engl. J. Med 372, 825–834 (2015). [PubMed: 25714160]
43. Deakin LJ, et al. The *Clostridium difficile spo0A* gene is a persistence and transmission factor. Infect. Immun 80, 2704–2711 (2012). [PubMed: 22615253]
44. Lewis BB, Pamer EG. Microbiota-based therapies for *Clostridium difficile* and antibiotic-resistant enteric Infections. Annu. Rev. Microbiol 71, 157–178 (2017). [PubMed: 28617651]
45. Abt MC, McKenney PT, Pamer EG. *Clostridium difficile* colitis: pathogenesis and host defence. Nat. Rev. Microbiol 14, 609–620 (2016). [PubMed: 27573580]

46. Sanchez-Romero MA, Cota I, Casadesus J. DNA methylation in bacteria: from the methyl group to the methylome. Curr. Opin. Microbiol 25, 9–16 (2015). [PubMed: 25818841]
47. Griffiths D, et al. Multilocus sequence typing of *Clostridium difficile*. J. Clin. Microbiol 48, 770–778 (2010). [PubMed: 20042623]
48. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012). [PubMed: 22388286]
49. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009). [PubMed: 19505943]
50. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010). [PubMed: 20110278]
51. Oliveira PH, Touchon M, Rocha EP. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. Nucleic Acids Res. 42, 10618–10631 (2014). [PubMed: 25120263]
52. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 43, D298–299 (2015). [PubMed: 25378308]
53. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30, 1575–1584 (2002). [PubMed: 11917018]
54. Katoh K, Standley DM. MAFFT: iterative refinement and additional methods. Methods Mol. Biol 1079, 131–146 (2014). [PubMed: 24170399]
55. 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). [PubMed: 19854763]
56. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39, W29–37 (2011). [PubMed: 21593126]
57. Bland C, et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinf. 8, 209 (2007).
58. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res. 31, 371–373 (2003). [PubMed: 12520025]
59. Goldfarb T, et al. BREX is a novel phage resistance system widespread in microbial genomes. EMBO J. 34, 169–183 (2015). [PubMed: 25452498]
60. Ofir G, et al. DISARM is a widespread bacterial defence system with broad anti-phage activities. Nat. Microbiol 3, 90–98 (2018). [PubMed: 29085076]
61. Makarova KS, Wolf YI, van der Oost J, Koonin EV. 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). [PubMed: 19706170]
62. Doron S, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. Science, (2018).
63. Xie Y, et al. TADB 2.0: an updated database of bacterial type II toxin-antitoxin loci. Nucleic Acids Res. 46, D749–D753 (2018). [PubMed: 29106666]
64. Fouts DE. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res. 34, 5839–5851 (2006). [PubMed: 17062630]
65. Arndt D, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 44, W16–21 (2016). [PubMed: 27141966]
66. Cury J, Jove T, Touchon M, Neron B, Rocha EP. Identification and analysis of integrons and cassette arrays in bacterial genomes. Nucleic Acids Res. 44, 4539–4550 (2016). [PubMed: 27130947]
67. Cury J, Touchon M, Rocha EPC. Integrative and conjugative elements and their hosts: composition, distribution and organization. Nucleic Acids Res. 45, 8943–8956 (2017). [PubMed: 28911112]
68. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113 (2004). [PubMed: 15318951]

69. 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). [PubMed: 20626897]
70. Stamatakis A RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313 (2014). [PubMed: 24451623]
71. Touchon M, et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet. 5, e1000344 (2009). [PubMed: 19165319]
72. Miele V, Penel S, Duret L. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12, 116 (2011). [PubMed: 21513511]
73. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr. Opin. Microbiol. 11, 472–477 (2008). [PubMed: 19086349]
74. Didelot X, Wilson DJ. ClonalFrame ML: efficient inference of recombination in whole bacterial genomes. PLoS Comput. Biol. 11, e1004041 (2015). [PubMed: 25675341]
75. Sawyer S Statistical tests for detecting gene conversion. Mol. Biol. Evol. 6, 526–538 (1989). [PubMed: 2677599]
76. Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290 (2004). [PubMed: 14734327]
77. Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol. Biol. Evol. 31, 1929–1936 (2014). [PubMed: 24739305]
78. Csuros M Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26, 1910–1912 (2010). [PubMed: 20551134]
79. Ng YK, 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). [PubMed: 23405251]
80. Sorg JA, Dineen SS. Laboratory maintenance of Clostridium difficile. Curr. Protoc. Microbiol Chapter 9, Unit9A 1 (2009).
81. Cartman ST, Minton NP. A mariner-based transposon system for in vivo random mutagenesis of *Clostridium difficile*. Appl. Environ. Microbiol 76, 1103–1109 (2010). [PubMed: 20023081]
82. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, 3rd, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods 6, 343–345 (2009). [PubMed: 19363495]
83. Donnelly ML, Li W, Li YQ, Hinkel L, Setlow P, Shen A. A Clostridium difficile-specific, gel-forming protein required for optimal spore germination. MBio 8, (2017).
84. Ducret A, Quardokus EM, Brun YV. MicrobeJ, a tool for high throughput bacterial cell detection and quantitative analysis. Nat. Microbiol 1, 16077 (2016). [PubMed: 27572972]
85. Shen A, Fimlaid KA, Pishdadian K. Inducing and quantifying *Clostridium difficile* spore formation. Methods Mol. Biol 1476, 129–142 (2016). [PubMed: 27507338]
86. Ribis JW, Ravichandran P, Putnam EE, Pishdadian K, Shen A. The conserved spore coat protein SpoVM Is largely dispensable in Clostridium difficile spore formation. mSphere 2, (2017).
87. Edwards AN, Karim ST, Pascual RA, Jowhar LM, Anderson SE, McBride SM. Chemical and stress resistances of Clostridium difficile spores and vegetative cells. Front. Microbiol 7, 1698 (2016). [PubMed: 27833595]
88. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE 5, e11147 (2010). [PubMed: 20593022]
89. Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT. Reordering contigs of draft genomes using the Mauve aligner. Bioinformatics 25, 2071–2073 (2009). [PubMed: 19515959]
90. Danecek P, et al. The variant call format and VCFtools. Bioinformatics 27, 2156–2158 (2011). [PubMed: 21653522]
91. Novichkov PS, et al. RegPrecise 3.0--a resource for genome-scale exploration of transcriptional regulation in bacteria. BMC Genomics 14, 745 (2013). [PubMed: 24175918]
92. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. Bioinformatics 14, 48–54 (1998). [PubMed: 9520501]

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). [PubMed: 24470570]
94. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014). [PubMed: 24695404]
95. Kopylova E, Noe L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28, 3211–3217 (2012). [PubMed: 23071270]
96. Quast C, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590–596 (2013). [PubMed: 23193283]
97. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucleic Acids Res. 31, 439–441 (2003). [PubMed: 12520045]
98. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009). [PubMed: 19451168]
99. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930 (2014). [PubMed: 24227677]
100. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014). [PubMed: 25516281]
101. Huang DW, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 35, W169–175 (2007). [PubMed: 17576678]
102. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676 (2005). [PubMed: 16081474]
103. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 47, D419–D426 (2019). [PubMed: 30407594]
104. Wang M, Zhao Y, Zhang B. Efficient test and visualization of multi-set intersections. Sci. Rep 5, 16923 (2015). [PubMed: 26603754]
105. Anjuwon-Foster BR, Maldonado-Vazquez N, Tamayo R. Characterization of flagellum and toxin phase variation in *Clostridioides difficile* ribotype 012 isolates. J. Bacteriol 200, (2018).
106. Chen X, et al. A mouse model of *Clostridium difficile*-associated disease. Gastroenterology 135, 1984–1992 (2008). [PubMed: 18848941]
107. McKee RW, Aleksanyan N, Garrett EM, Tamayo R. Type IV pili promote *Clostridium difficile* adherence and persistence in a mouse model of infection. Infect. Immun 86, (2018).
108. Woods EC, Edwards AN, Childress KO, Jones JB, McBride SM. The C. difficile clnRAB operon initiates adaptations to the host environment in response to LL-37. PLoS Pathog. 14, e1007153 (2018). [PubMed: 30125334]
109. Purcell EB, et al. A nutrient-regulated cyclic diguanylate phosphodiesterase controls *Clostridium difficile* biofilm and toxin production during stationary phase. Infect. Immun 85, (2017).
110. Pereira FC, et al. The spore differentiation pathway in the enteric pathogen *Clostridium difficile*. PLoS Genet. 9, e1003782 (2013). [PubMed: 24098139]
111. Serrano M, et al. A recombination directionality factor controls the cell type-specific activation of sigmaK and the fidelity of spore development in *Clostridium difficile*. PLoS Genet. 12, e1006312 (2016). [PubMed: 27631621]
112. Theriot CM, Koumpouras CC, Carlson PE, Bergin, II, Aronoff DM, Young VB. Cefoperazone-treated mice as an experimental platform to assess differential virulence of *Clostridium difficile* strains. Gut Microbes 2, 326–334 (2011). [PubMed: 22198617]

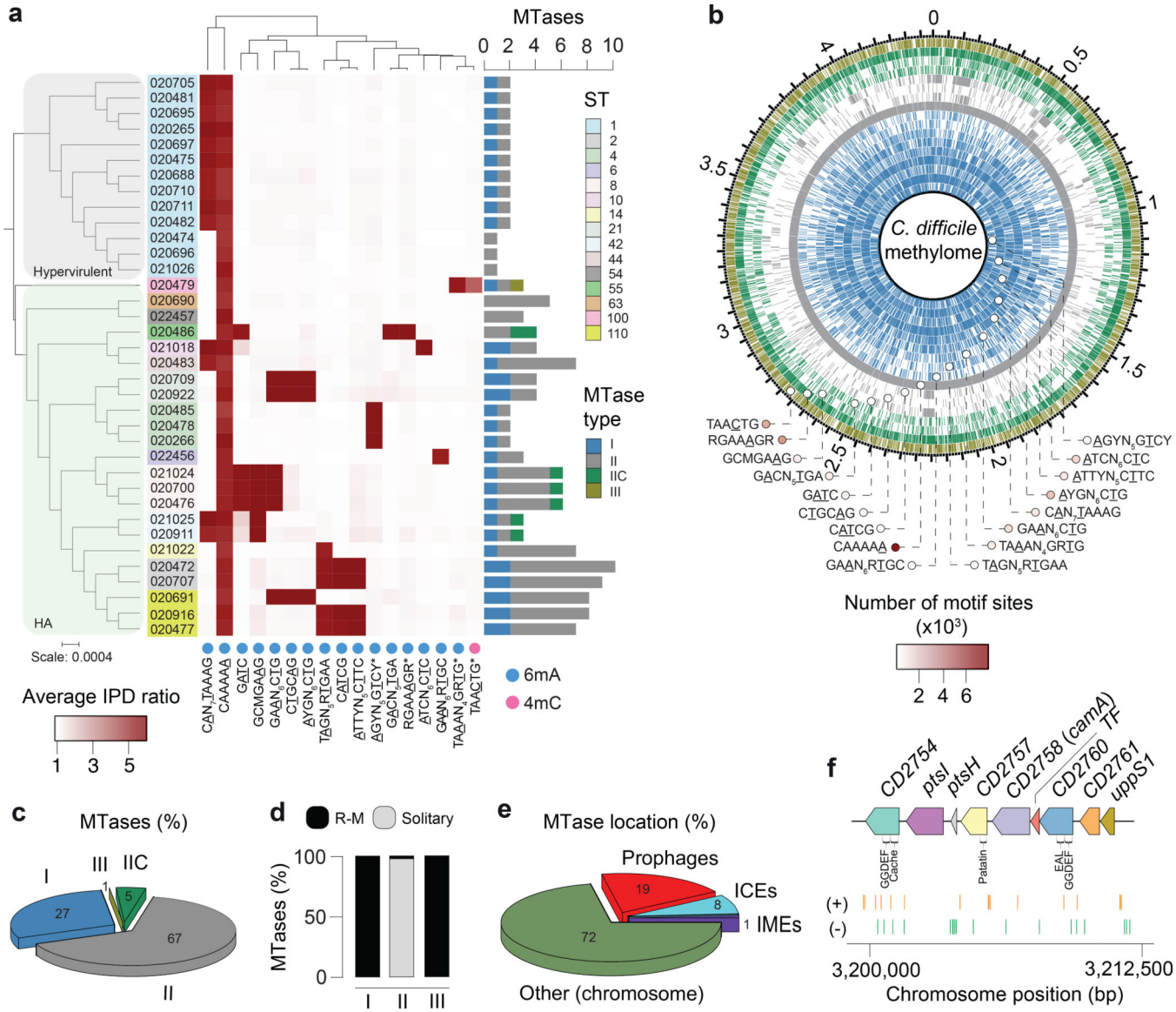


Fig. 1.

Methylomes of the 36 *C. difficile* strains. (a) Phylogenetic tree of the 36 *C. difficile* strains colored by clade (hypervirulent, human and animal (HA) associated) and MLST sequence type (ST). Heatmap depicting the landscape of methylated motifs per genome, and their average interpulse duration (IPD) ratio. Asterisks refer to new motifs not previously listed in the reference database REBASE. Methylated bases are underlined. The CAAAAA motif was consistently methylated across isolates. Barplot indicates the number and types of active MTases detected per genome. In Type IIC systems, MTase and REase are encoded in the same polypeptide. (b) Representation of the *C. difficile* methylome. Shown are the positions of all methylation motif sites in the reference genome of *C. difficile* 630, colored according to MTase type. Also shown are the average motif occurrences per genome (across the 36 isolates). (c) % of MTases detected according to type. (d) % 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. + / – signs correspond to the sense and antisense strands respectively. Vertical bars correspond to the distribution of the CAAAAA motif. *CD2754*: phosphodiesterase with a GGDEF domain (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; *CD2762*: undecaprenyl diphosphate synthase. The genomic context of *camA* is largely conserved across strains, located ~25 kb upstream of the S-layer biogenesis locus (Extended Data Figs. 4c,d). Several of the genes flanking *camA* (including itself) are part of the *C. difficile* core-genome (see below), suggesting that they may play biological roles fundamental to *C. difficile*.

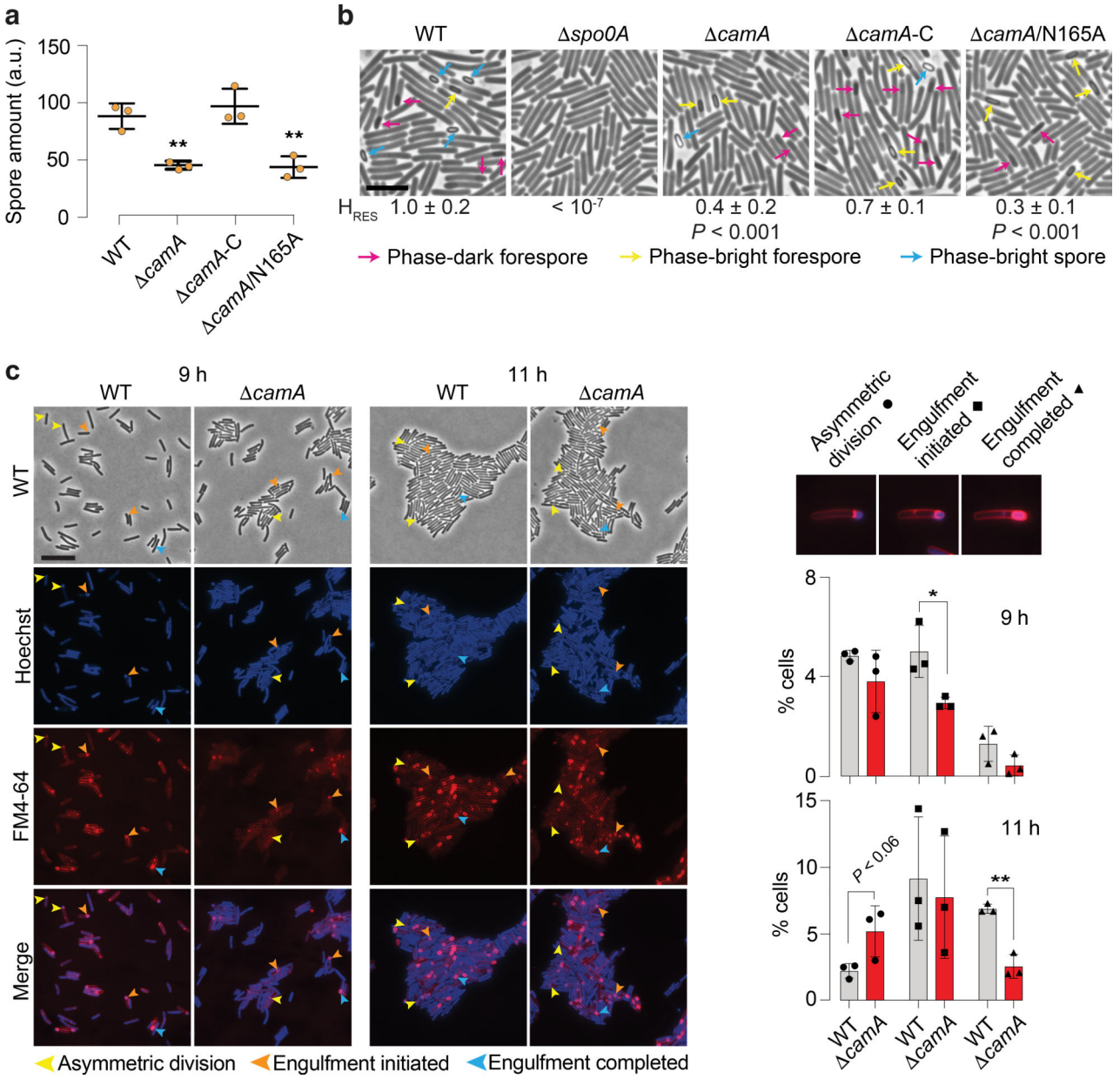


Fig. 2.

CamA modulates sporulation levels in *C. difficile*. (a) Spore purification efficiencies obtained from sporulating cells ($n = 3$ independent spore preparations, $^{**}P < 10^{-2}$; one-way ANOVA and Tukey's test). The spore yield (arbitrary units, a.u.), was determined by measuring the optical density at 600 nm of the resulting spore preparations and correcting for the volume of re-suspension water. Data are presented as mean ± standard deviation (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[43]. Immature phase-dark forespores are marked in pink, and mature phase-bright forespores and free spores are shown in yellow and blue, respectively. Scale bar represents 5 μm. Heat resistance (H_{RES}) efficiency values are also provided as mean ± standard deviation of $n = 3$ independent

replicates. (c) Morphological analysis of wild-type and Δ*camA* cells using fluorescent stains comparing 9 and 11 h following sporulation induction. The polar septum formed during asymmetric division is visible using FM4–64 membrane staining, while 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 also allows the engulfing membranes to be visualized. 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 mother cell cytosol, and both stains are excluded. Yellow arrows show cells that are undergoing asymmetric division (indicated by a flat polar septum); orange arrows show cells that are in the process of engulfment (indicated by a curved polar septum); and blue arrows show cells that have completed engulfment (indicated by bright membrane staining fully surrounding the forespore). Scale bar: 10 μm. Barplots indicate the percentage of sporulating cells at different stages of spore assembly in both wild-type and Δ*camA* cells. Data is presented as mean ± standard deviation of $n = 3$ independent replicates. 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. * $P \leq 0.05$, ** $P < 10^{-2}$, two-tailed unpaired Student's t-test.

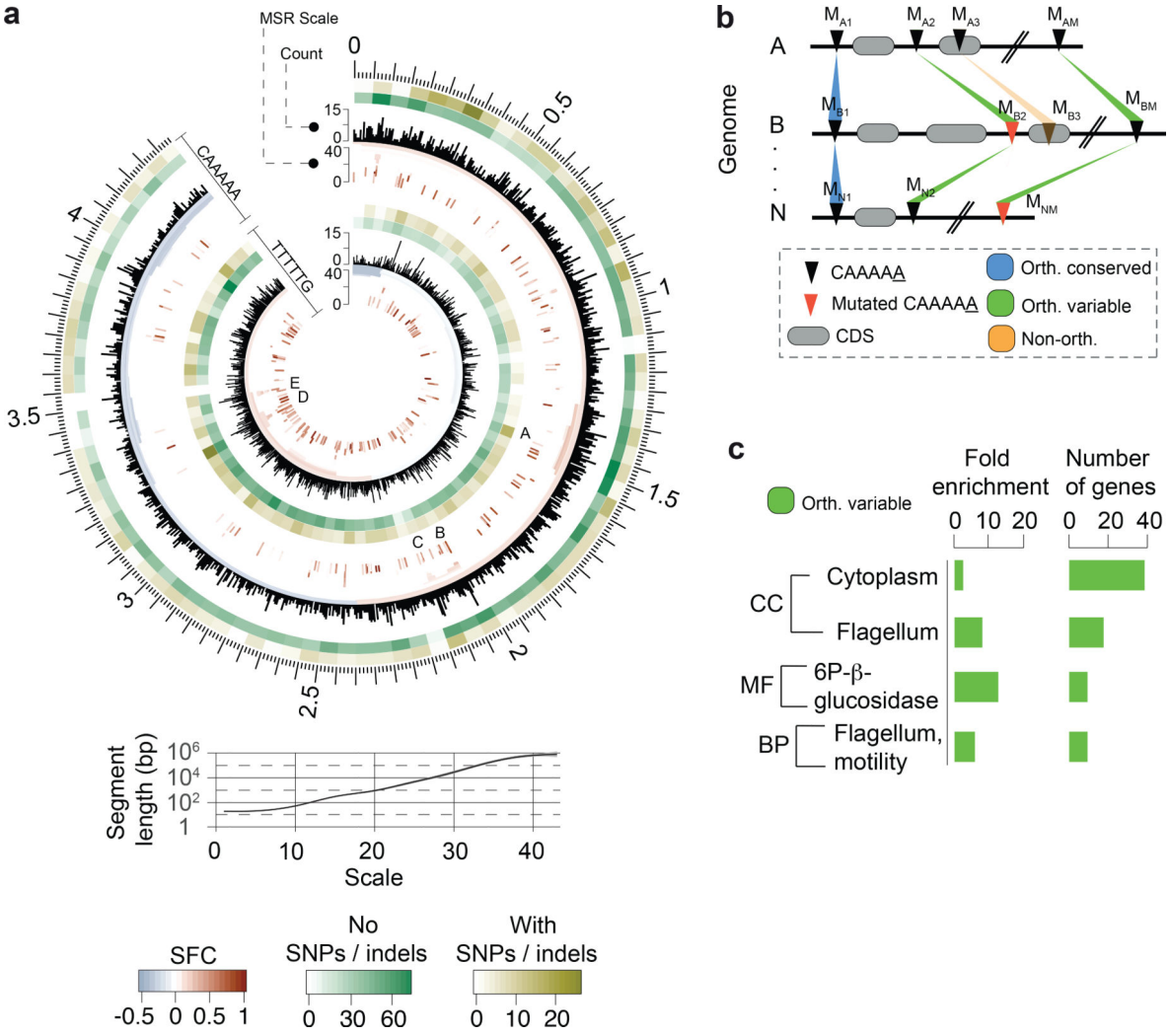


Fig. 3.

Abundance, distribution, and conservation of CAAAAA motif sites. (a) Representation of distribution of CAAAAA sites in both strands of the reference *C. difficile* 630 genome and corresponding genomic signal obtained by multi-scale signal representation (MSR). Briefly, 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. While scale values <10 are typically associated with regions <100 bp, genomic regions enriched for CAAAAA sites at scale values >20 correspond to segments larger than 1 kb (*i.e.*, gene and operon scale). Letters (A-E) represent regions with particularly high abundance of CAAAAA motif sites, including genes related to sporulation (e.g., *spo0A*, *spoIIIAA-AH*, *spoIVB*, *sigK*), membrane transport (PTS and ABC-type systems), transcriptional regulation (*e.g.*, *iscR*, *fur*), and coding for multiple cell wall proteins (Supplementary Table 6d). Relation between MSR scale and segment length is also shown. The significant fold-change (SFC) corresponds to the fold-change ($\log_2$ ratio) between observed and randomly expected overlap statistically significant at $P = 10^{-6}$ based on the Z-test. Heatmap layers correspond to the number of orthologous conserved (no SNPs/indels, green-shaded) and orthologous

variable (with SNPs/indels) CA_5 motif positions. (b) Whole genome alignment of 37 *C. difficile* genomes (36 isolates + *C. difficile* 630 as reference) was performed using Mauve. We defined an orthologous occurrence of the CAAAAA motif (black triangles) if an exact match to the motif was present in each of the 37 genomes (conserved, blue-shaded regions), 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-shaded regions). Non-orthologous CAAAAA positions are indicated as orange-shaded regions. The results are shown in Fig. 3a in the form of heatmaps. Numbering in 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 found to over-represent orthologous variable CAAAAA positions include cytoplasm- (e.g., *pheA*, *fdhD*, *ogt1*, *spoIVA*) and motility-related genes (e.g., *fliZ*, *fliN*, *fliM*, *flgL*). Single categories were considered significantly enriched at $P < 0.05$ (one-tailed Fisher's exact test, FDR corrected) and correspond to 73 out of a total of 617 genes analyzed.

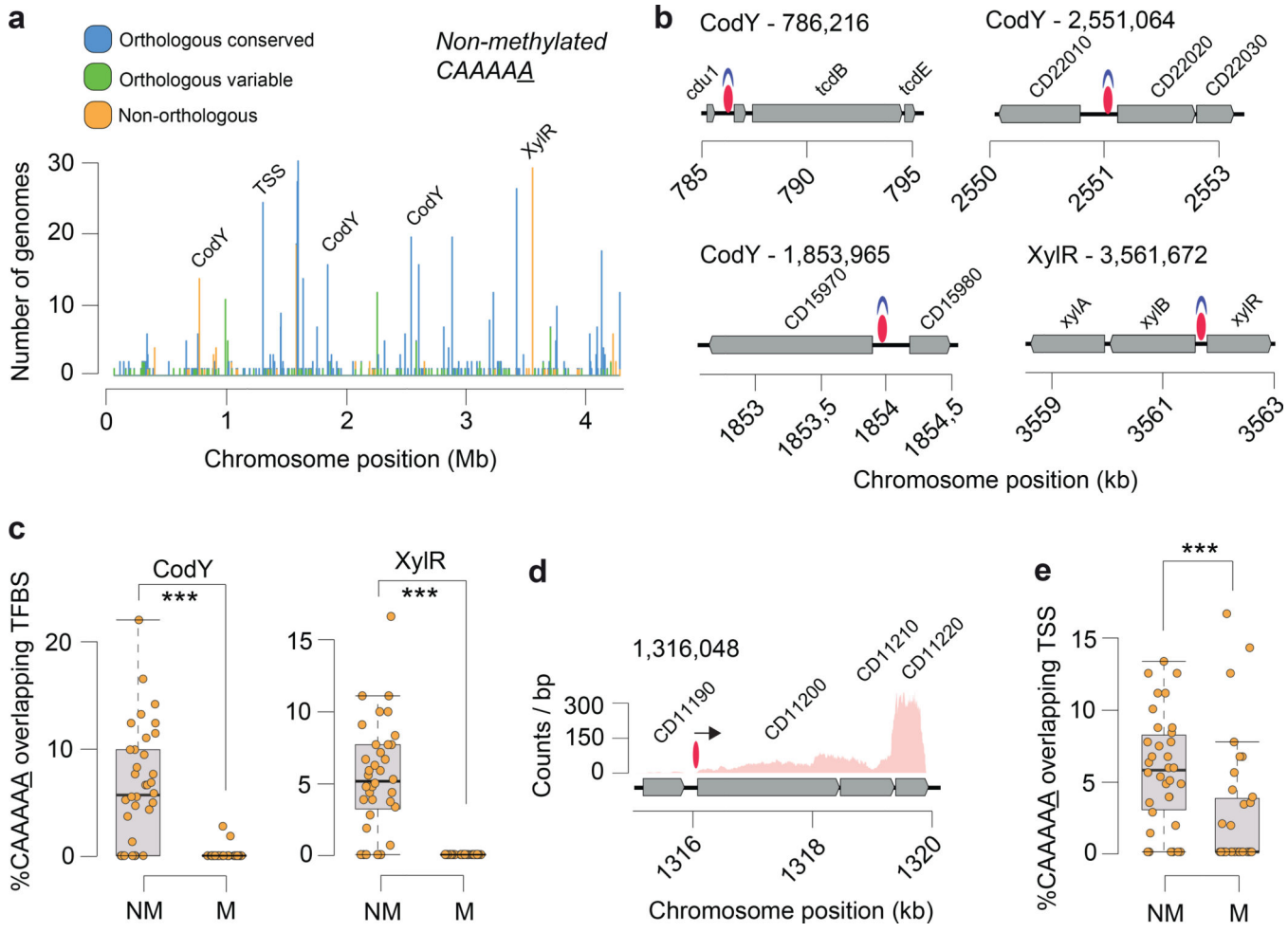


Fig. 4.

Distribution of non-methylated CAAAAA motif sites, and overlap with transcription factor binding sites (TFBS) and transcription start sites (TSS). (a) Number of *C. difficile* isolates for 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 colors correspond to orthologous (conserved and variant) and non-orthologous CAAAAA positions. Some of the major peaks of non-methylated CAAAAA positions were found to overlap with TFBS (*e.g.*, CodY, XylR) and TSS. (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 illustrated in Extended Data Fig. 7b. (c) % CAAAAA motif sites (non-methylated and methylated) overlapping CodY and XylR for each of the $n = 36$ *C. difficile* isolates. (d) Example of a chromosomal region in which non-methylated CAAAAA motifs overlap a TSS (shown as arrow). (e) % CAAAAA motifs (non-methylated (NM) and methylated (M)) overlapping TSSs for each of the $n = 36$ *C. difficile* isolates. 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. ***$P < 10^{-3}$ (one-sided Mann-Whitney-Wilcoxon rank sum test with continuity correction).

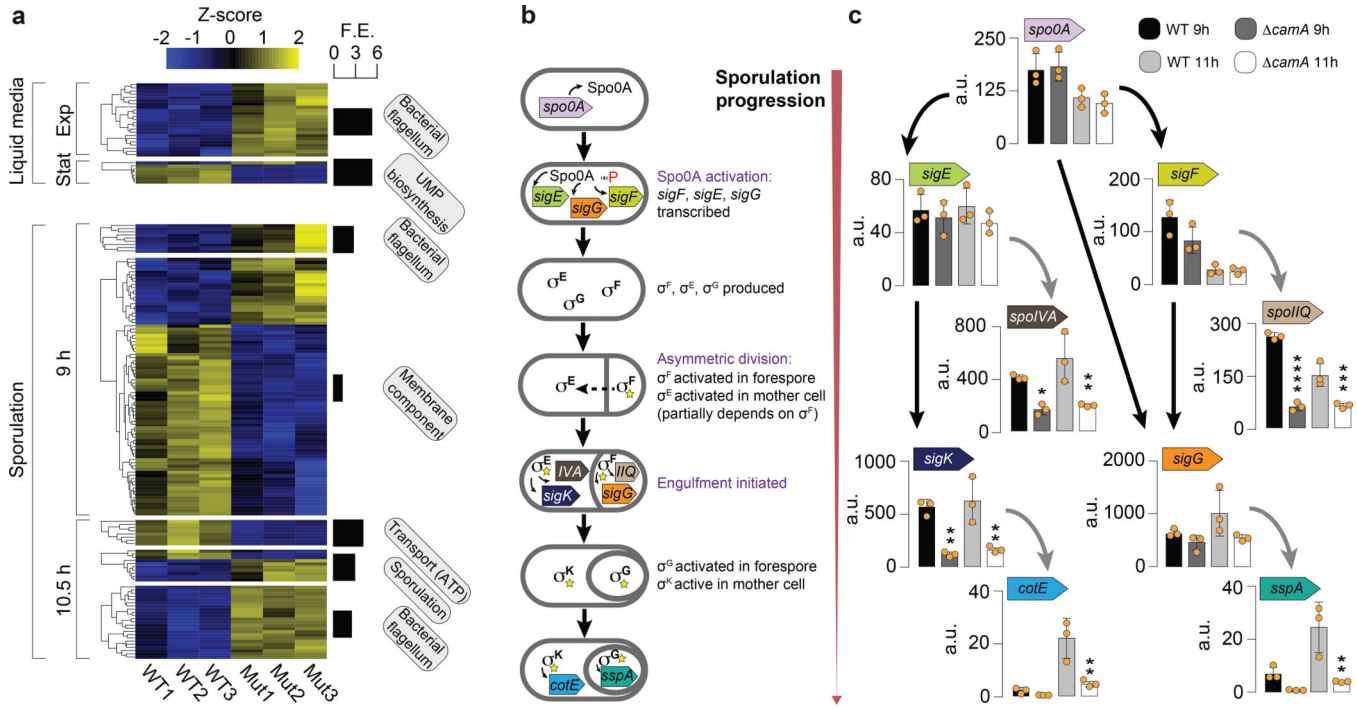


Fig. 5.

Gene expression analysis. (a) Heatmap of 161 genes in three replicates of *C. difficile* 630 compared to equal number of replicates of *C. difficile* 630Δ*camA* and that are enriched for the GO terms shown in boxes and detailed in Supplementary Table 8c. The *Z* score reflects the degree of down- (*Z* score < 0) or up- (*Z* score > 0) regulation, computed by subtracting the mean of the log-transformed expression values and dividing by the standard deviation for each gene over all samples scored. (b) Schematic illustrating 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 only becomes active in the forespore after asymmetric division is completed[110]. Activated σ^F subsequently induces the transcription of genes whose products mediate σ^G activation in the forespore and partially mediates σ^E activation in the mother cell[34]. Activated σ^E induces the transcription of *sigK*[32] and factors required for the excision of a prophage-like element from the *sigK* gene[111]. Thus, *C. difficile* sporulation is 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) Comparison of relative transcript levels in wild type and Δ*camA* as determined by qRT-PCR for sporulation sigma factor genes and representative genes in the regulons of sporulation-specific sigma factors at 9 and 11 h after sporulation induction (a separate set of $n = 3$ RNA sample replicates was used). It should be noted that the primers for *sigK* amplify a region prior to the *sigK* excision site[111]. Data is presented as mean ± standard deviation. Statistical significance was determined by one-way

ANOVA and Tukey's test for multiple comparisons (* $P \leq 0.05$, ** $P < 10^{-2}$, *** $P < 10^{-3}$, **** $P < 10^{-4}$).

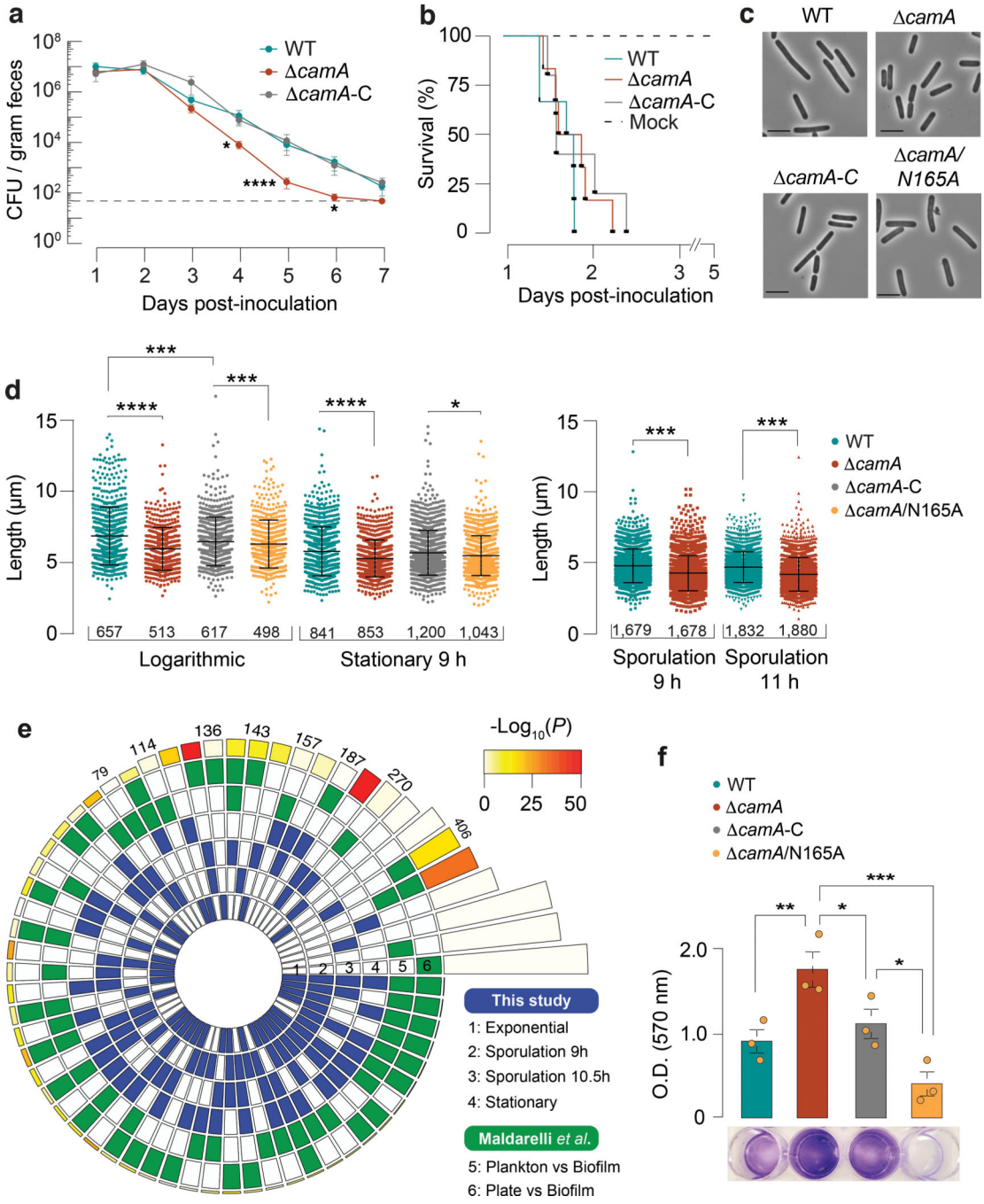


Fig. 6.

In vivo and additional functional impacts of the Δ*camA* mutation. (a) Kinetics of infection in antibiotic-treated mice ($n = 12$) following inoculation with a sub-lethal amount (10^5 spores) of wild type (WT) *C. difficile* 630Δ*erm*, MTase mutant Δ*camA*, and complement Δ*camA-C*. When inoculated with 630Δ*erm* strains, antibiotic-treated mice typically do not develop fulminant disease and instead serve as a model of intestinal colonization and persistence by *C. difficile*[106,107,112]. Dotted line indicates the limit of detection. Data are presented as mean ± standard error of the mean. Log_{10}-transformed data from each time point were analyzed

by ANOVA for each time point. * $P < 0.05$, **** $P < 10^{-4}$. (b) Kaplan-Meier survival curves for clindamycin-treated golden Syrian hamsters ($n = 6$) infected with 10^3 spores of either wild type (WT) *C. difficile* 630Δ*erm*, Δ*camA*, and 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* (scale bar 5 μm). (d) Comparison of cell length. Data are presented as mean ± standard deviation of $n = 3$ independent biological replicates (exact cell numbers measured are given in the figure). * $P < 0.05$, *** $P < 10^{-3}$, **** $P < 10^{-4}$ (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-shaded, $n = 1{,}537$) and those from Maldarelli *et al.*[39] (green-shaded, $n = 1{,}735$). The latter corresponds to *C. difficile* DE genes in conditions favoring biofilm formation compared to growth on a plate or planktonic form. Color 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 to DE genes in conditions favoring the production of biofilm ($P < 10^{-9}$, one-tailed hypergeometric test implemented in *SuperExactTest*, Bonferroni adjusted). (f) Biofilm production as 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 latter retained some DNA binding ability capable of altering transcription of some genes even in the absence of methylation. Data are presented as mean ± standard deviation of $n = 3$ independent biological replicates, with each strain assayed in quadruplicate in each experiment. * $P < 0.05$, ** $P < 10^{-2}$, *** $P < 10^{-3}$, two-way ANOVA with Dunnett's post-test.