Epigenomic landscape of the human pathogen *Clostridium difficile*

Pedro H. Oliveira[1], Alex Kim[1], Ognjen Sekulovic[2], Elizabeth M. Garrett[3], Dominika Trzilova[3], Edward A. Mead[1], Theodore Pak[1], Shijia Zhu[1], Gintaras Deikus[1], Marie Touchon[4,5], Colleen Beckford[1], Nathalie E. Zeitouni[1], Deena Altman[1,6], Elizabeth Webster[1], Irina Oussenko[1], Aneel K. Aggarwal[7], Ali Bashir[1], Gopi Patel[6], Camille Hamula[6], Shirish Huprikar[6], Richard J. Roberts[8], 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, Infectious Disease, 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

[8] New England Biolabs, Ipswich, Massachusetts, United States of America

[#] To whom correspondence should be addressed. Gang Fang (gang.fang@mssm.edu)

Keywords: DNA methylation; methyltransferase; SMRT sequencing; sporulation; restriction-modification systems.

Abstract

Clostridium 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 explored. Here, we performed the first DNA methylome analysis of *C. difficile* using 36 human isolates and observed great epigenomic diversity. Strikingly, we discovered a DNA methyltransferase with a well-defined specificity, highly conserved across our dataset and in all the ~300 global *C. difficile* genomes we further examined. Inactivation of the methyltransferase negatively impacted sporulation, a key step in *C. difficile* transmission, consistently supported by multi-omics data and genetic experiments and a mouse infection model. Transcriptomic analysis also suggested that epigenetic regulation is associated with host colonization and biofilm formation. The epigenomic landscape also allowed an integrative analysis of multiple defense systems with respect to their roles in host defense and in regulating gene flux in *C. difficile*. These findings open up a new 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.

Introduction

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

Despite the significant progress achieved in the understanding of *C. difficile* physiology, genetics, and genomic evolution[9,10], the roles played by epigenetic factors, namely DNA methylation, have not been studied. In the bacterial kingdom, there are three major forms of DNA methylation: N6-methyladenine (m6A, the most prevalent form representing ~80%), N4-methylcytosine (m4C), and 5-methylcytosine (m5C). Although bacterial DNA methylation is most commonly associated with restriction-modification (R-M) systems that defend hosts against invading foreign DNA[11,12], increasing evidence suggests that DNA methylation also regulates a number of biological processes such as DNA replication and repair, cell cycle,

chromosome segregation and gene expression, among others[13-17]. 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)[18], which can detect all three types of DNA methylation, albeit at different signal-to-noise ratios: high for m6A, medium for m4C, and low for m5C. This technique enabled to characterize the first bacterial methylomes[19,20], and since then, more than 1,500 (as of 4/2018) have been mapped, heralding a new era of "bacterial epigenomics"[21].

Herein, we mapped and characterized 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, consistently supported by multi-omics data and genetic experiments. Further integrative transcriptomic analysis suggested that epigenetic regulation by DNA methylation is also associated with *C. difficile* host colonization and biofilm formation. Finally, the epigenomic landscape of *C. difficile* also allowed us to perform a data-driven joint analysis of multiple defense systems and their contribution to gene flux. 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 Project 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)[22] 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 (Materials and Methods). 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 m6A type, one motif (TAA<u>C</u>TG) belonged to the m4C type, and no confident m5C motifs were detected (Supplementary Text). Like most bacterial methylomes, >95% of the m6A and m4C 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 m6A motif, CAAAA<u>A</u>, was intriguingly present across all genomes, which led us to raise the hypothesis that m6A methylation events at this motif, and its corresponding MTase, may play an important and conserved functional role in *C. difficile*.

A DNA methyltransferase and its target motif are ubiquitous in *C. difficile*

Motivated by the consistent presence of the methylation motif CAAAA<u>A</u> across all the *C. difficile* isolates, we proceeded to examine the encoded MTases. From the 36 genomes assemblies, we identified a total of 139 MTases (average of 3.9 per genome) (Fig. 1a,

Supplementary Table 2b) representing all the four major types[23]. 38 MTases (27%) belong to Type I R-M systems, and 100 MTases (72%) belong to Type II (Fig. 1c). All but one of the Type II MTases are solitary, *i.e.*, devoid of a cognate restriction endonuclease (REase) (Fig. 1d). The least abundant MTases (1%) belong to Type III R-M systems (Fig. 1c). A Type IV (no cognate MTase) McrBC REase was also present in all genomes (Supplementary Table 2b). 28% of MTases were located in mobile genetic elements (MGEs): (19% in prophages and 9% in integrative conjugative/mobile elements), while the large majority was encoded in other chromosomal regions (Supplementary Tables 2b-d) (Fig. 1e). No MTases were found in plasmids.

Consistent with the presence of a highly conserved CAAAAA motif, we identified a Type II m6A solitary DNA MTase (577 aa) present across isolates (Fig. 1f, Supplementary Table 2b). This MTase is encoded by *CD2758* in *C. difficile* 630, a reference strain that was isolated from a *C. difficile* outbreak in Switzerland[9,24]. The ubiquity of this MTase 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 3). REBASE also showed functional orthologs of CD2758 only in very few other *Clostridiales* and *Fusobacteriales* (Supplementary Figs. 1a, b), suggesting CD2758 is fairly unique to *C. difficile*. The genomic context of *CD2758* is largely conserved across strains (Supplementary Fig. 1c), located ~25 kb upstream of the S-layer biogenesis locus (Supplementary Fig. 1d). Several of the genes flanking *CD2758* (including itself), are part of the *C. difficile* core-genome (see below), suggesting that they may play biological roles fundamental to *C. difficile*. Hence, CD2758 is a solitary Type II m6A MTase that specifically recognizes the CAAAAA motif, and is ubiquitous, and fairly unique to *C. difficile*.

Inactivation of *CD2758* reduces sporulation levels both *in vitro* and *in vivo*

To discover possible functional roles of CD2758, we searched among published transposon sequencing (Tn-Seq) studies of *C. difficile*. From a recent study analyzing *C. difficile* gene

essentiality and sporulation via Tn-Seq, we found a homolog of *CD2758 (R20291_2646)* among ~800 genes whose mutation impacted sporulation in the 027 isolate, R20291[25]. Given the critical role of sporulation in the persistence and dissemination of *C. difficile* in humans and hospital settings, we decided to test if *CD2758* inactivation could reduce spore purification efficiencies in the 630Δ*erm* strain background as seen for *R20291_26460* transposon mutants. We used allele-coupled exchange[26] to construct an in-frame deletion in 630Δ*erm CD2758* and to complement Δ*CD2758* with either wild-type *CD2758* or a variant encoding a catalytic site mutation (N165A) of the MTase in single copy from the *pyrE* locus (Materials and Methods; Supplementary Fig. 2a; Supplementary Tables 4a, b). We observed that spore purification efficiencies decreased by ~50% in the mutant relative to wild-type (Fig. 2a, $P < 10^{-2}$, ANOVA, Tukey's test). Complementation of Δ*CD2758* with the wild-type, but not the catalytic mutant construct, restored spore purification efficiencies to values similar to those observed in wild-type cells (Fig. 2a, Supplementary Table 4c). Hence, this complementation experiment supports that the loss of methylation events by CD2758, rather than the loss of non-catalytic roles of CD2758, led to the decrease in purified spores.

The decreased spore purification efficiencies observed in the Δ*CD2758* mutants could be due to reduced sporulation or defective spore formation[27]. Visual inspection of samples before and after spore purification on a density gradient revealed qualitatively lower levels of phase-bright spores (Supplementary Fig. 2b). Since purified wild-type and Δ*CD2758* spores had similar levels of chloroform resistance (Supplementary Fig. 2c), and wild-type and Δ*CD2758* spores germinated with similar efficiency when plated on media containing germinant (Supplementary Fig. 2b), the reduced spore purification efficiencies of the MTase mutants likely reflect a defect in sporulation initiation rather than the sporulation process itself. Consistent with this hypothesis, sporulation levels were significantly reduced in the Δ*CD2758* mutant and the catalytic mutant complementation strain by ~50% relative to wild-type and the wild-type complementation strain based on heat resistance (Fig. 2b, $P < 10^{-3}$,

ANOVA, Tukey's test) and phase contrast microscopy analyses of sporulating cultures (Fig. 2c).

The effect of the Δ*CD2758* mutation was further evaluated in a mouse model[28]. In brief, C57BL/6 mice were administered a cocktail of antibiotics in their water for three days, then a single intra-peritoneal dose of clindamycin 2 days prior to inoculation. Mice were inoculated with 10^5 spores by oral gavage. Six mice were used for each of the three genotypes: wild-type, Δ*CD2758*, and Δ*CD2758-C* (complementation of Δ*CD2758* with the MTase). No mortality was observed at the given doses of *C. difficile* spores. Fecal samples were collected every 24 hours for seven days. 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. As expected, CFU levels decreased steadily from day 2 post-inoculation to day 7 (Fig. 2d). Notably, the Δ*CD2758* mutant showed CFU levels 10-100 times lower than those observed in the wild-type and complement strains. The bacteria were cleared from the feces 6 days post-inoculation for the MTase mutant, while they were still detectable at days 6 and 7 for the wild-type and complement strains.

Taken together, these findings suggest that CAAAAA methylation events play an important role in sporulation both *in vitro* and *in vivo*, and this sporulation defect is more pronounced in the latter. These observed functional differences prompted us to perform a comprehensive methylome and transcriptome analysis of the wild-type and MTase mutant strains.

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

The *C. difficile* genome has an average of 7,721 (±197, sd) CAAAAA motif sites (Supplementary Table 5a). Adjusted by the k-mer frequency of the AT-rich *C. difficile* genome (70.9%) using Markov models[29] (Materials and Methods), CAAAAA motif sites are significantly under-represented in the chromosome, particularly in intragenic regions

(Supplementary Fig. 3a, Supplementary Table 5b). To evaluate if specific chromosomal regions are enriched or depleted for this motif, we used a multi-scale signal representation (MSR) approach[30] (Materials and Methods; Fig. 3a). 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, genomics regions enriched for CAAAAA sites at scale values >20 correspond to segments larger than 1 kb (*i.e.* gene and operon scale), and include genes related to sporulation and colonization (Fig. 3a, regions A-E; Supplementary Tables 5c, d): stage 0 (*spo0A*), stage III (*spoIIIAA-AH*), and stage IV (*spoIVB*, *sigK*) of sporulation, membrane transport (PTS and ABC-type transport systems), transcriptional regulation (*e.g*. *iscR*, *fur*), and multiple cell wall proteins (CWPs).

To further characterize CAAAAA motif sites, we built on the large collection of methylomes of the same species and sought to categorize these motif sites on the basis of their positional conservation across genomes. We performed whole genome alignment of 37 *C. difficile* genomes (36 isolates + *C. difficile* 630 as reference) (Materials and Methods), 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). We found a total of 5,828 conserved orthologous motif positions, 1,050 variable orthologous positions (885 with SNPs and 165 with indels), and an average of 843 non-orthologous positions per genome (Supplementary Table 5e). The latter were, as expected, largely mapped to MGEs. Among orthologous positions, the variable ones represent a particularly interesting subset to study, since they contribute to variations across genomes of CAAAAA sites with subsequent methylation abrogation. We used DAVID functional analysis and found cytoplasm- and motility-related genes to over represent orthologous variable CAAAAA positions (FDR < 5%, Fig. 3c). Concomitantly, the regions with the highest density of orthologous variable positions were found at the S-layer locus region (region D, Fig. 3a), and between positions 0.31-0.33 Mb (Fig. 3a), rich in

flagellar genes. The very large number and dispersion of conserved orthologous positions precluded a similar functional analysis. To test whether homologous recombination (HR) contributes to the cross-genome variation of CAAAAA motif sites located in the core-genome, we also performed a systematic analysis of such events (Supplementary Text, Supplementary Figs. 3b-d) and found that HR tracts indeed over-represent (O/E=1.40, $P < 10^{-3}$; Chi-square test) orthologous variable CAAAAA motif positions, while the core-genome without HR tracts underrepresents them (O/E=0.89, $P < 10^{-3}$; Chi-square test) (Supplementary Figs. 3e, f).

Collectively, genome-wide distribution analyses and across-genome comparative analyses suggest that CAAAAA sites are enriched in regions harboring genes related to sporulation and colonization, orthologous variable CAAAAA positions are enriched in regions harboring genes related to cytoplasm- and motility-related genes, and the former's variability is at least partially fueled by HR.

Non-methylated CAAAAA motif sites are enriched in regulatory elements

DNA methylation is highly motif driven in bacteria; i.e. in most cases, >95% of the occurrence of a methylation motif is methylated. However, a small fraction of methylation motif sites can be non-methylated. The on/off switch of DNA methylation in a bacterial cell can contribute to epigenetic regulation through competitive binding between DNA MTases and other DNA binding proteins (e.g. transcription factors, TFs) as previously described for *E. coli* [31,32]. Previous bacterial methylome studies analyzing one or few genomes usually had insufficient statistical power to perform a systematic interrogation of non-methylated motifs sites. Building on the rich collection of 36 *C. difficile* methylomes, we performed a systematic detection and analysis of non-methylated CAAAAA sites. To classify a CAAAAA site as non-methylated, we adopted stringent filtering criteria at interpulse duration ratio (ipdR) and sequencing coverage, and found an average of 21.5 non-methylated CAAAAA motif sites per genome (Supplementary Fig. 4a; Supplementary Table 6a). Non-methylated motif sites were found dispersed throughout the full length of the *C. difficile* genome, 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) ($P < 10^{-4}$; 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 (Fig. 4a, Supplementary Table 6a), 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), e.g. upstream of the pathogenicity locus (PaLoc) (position 786,216), in prophage genes (1,593,616), within the *atp* operon (3,430,190), and in the xylose operon (3,561,672) (Supplementary Fig. 4b).

The non-methylated CAAAAA positions detected across the 36 *C. difficile* genomes allowed a systematic search for evidence of overlap with TF binding sites (TFBSs) and transcription start sites (TSSs). To test this, we queried our genomes for putative binding sites of 21 TFs pertaining to 14 distinct families (Supplementary Table 6b; Materials and Methods) and found overlaps between prominent peaks of non-methylated CAAAAA positions (Fig. 4a) and the TFBSs of CodY and XylR (Fig. 4b, Supplementary Fig. 4b, Supplementary Table 6c). 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; Supplementary Fig. 4c). In a similar enrichment analysis using 2,015 TSSs reconstructed from RNA-seq data coverage[33] (Materials and Methods; Supplementary Table 6d), we found a genome-level enrichment: of non-methylated CAAAAA sites preferentially overlapped with TSSs (Figs. 4d, e; Supplementary Figs. 4d, e; $P < 10^{-3}$, Mann-Whitney-Wilcoxon test,).

In addition to the on/off epigenetic switch driven by competitive binding between the MTase and other DNA binding proteins at CAAAAA sites, we hypothesized that epigenetic

heterogeneity within a clonal population could also stem from DNA replication errors at CAAAAA sites, especially because homopolymer tracts are expected to be more error-prone. To test this hypothesis, we investigated if CAAAAA had a larger than expected frequency of mutated reads (compared to baseline sequencing errors). Specifically, we tested if CAAAAA sites in *C. difficile* are indeed particularly error prone during DNA replication, compared to the control motifs TAAAAA or GAAAAA (here called KAAAAA). To account for the confounding effect of mutation rate at different sequence contexts, we further computed the % mutated reads (SNPs + indels) in four distinct species with a broad dispersion of mutation rates (*Mycobacterium tuberculosis* < *E. coli* ≈ *C. difficile* < *Helicobacter pylori*). Interestingly, we indeed found a significantly higher % of mutated reads mapping to CAAAAA sites compared to KAAAAA sites in *C. difficile* (Fig. 4f, Supplementary Fig. 4f), while this difference was not observed in any of the control genomes.

Collectively, these results highlight two types of variations that affect CAAAAA methylation status in *C. difficile*: on/off epigenetic switch of CAAAAA sites preferentially overlap with putative TFBSs and TSSs, and higher mutation rates at the CAAAAA motif sites that contribute to cell-to-cell heterogeneity.

Loss of CAAAAA methylation impacts transcription of sporulation genes

To study the functional significance of methylation at CAAAAA sites, we used RNA-seq to compare the transcriptomes of wild-type *C. difficile* 630Δ*erm* with that of the knockout mutant Δ*CD2758*: during exponential (mid-log) growth phase, following sporulation induction (15 and 19 h), and during stationary phase (Materials and Methods, Supplementary Table 7a, Supplementary Fig. 5). Of the 3,896 genes annotated in *C. difficile* 630, 405 – 715 (10.4 – 18.3%, depending on the growth stage) were differentially expressed (DE) at a 1% FDR (Supplementary Table 7b), with effect sizes ranging from $-3.7 \leq \log_2 FC \leq 3.0$ (between 13-fold for under-expressed and 8-fold for over-expressed genes). The set of DE genes was

enriched for the sporulation Gene Ontology (GO) term, as well as some additional terms: flagellum, cell division, ribosome/translation, ATP-coupled transport, and de-novo UMP biosynthesis (all $P < 10^{-3}$ and FDR < 5%) (Fig. 5a, Supplementary Table 7c). For validation purposes, qRT-PCR was performed on a selection of DE genes from Fig. 5a at different time points. The results largely validated the RNA-Seq analyses (Supplementary Fig. 6a).

Intriguingly, Fig. 5a shows the upregulation of several sporulation-related genes in Δ*CD2758* compared to wild-type, which first seemed to be in apparent conflict with the decrease in spore formation observed in Fig. 2a. To potentially resolve these observations and obtain a better understanding of CD2758's involvement in sporulation, we generated a heatmap restricted to DE genes induced during sporulation and under the control of the sporulation-specific sigma factors (SigF, SigE, SigG, and SigK) and master transcriptional activator of sporulation Spo0A (Supplementary Fig. 6b, Supplementary Table 7d)[34,35]. Sporulation gene expression is often divided into early and late based on the observation that the late-stage sigma factors, SigG and SigK, require the early-stage sigma factors, SigF and SigE, for activity in the forespore and mother cell compartments, respectively[36]. At the 15 h time point, both early and late stage sporulation-specific genes were under-expressed in the mutant relative to the wild-type, although early-acting Spo0A-regulated genes, and a small subset of SigE-regulated genes, were over-expressed in the mutant. By the 19 h time point, essentially only late-stage SigK-dependent genes were under-expressed in the mutant relative to the wild-type. Given the global impact of *CD2758* mutation on sporulation gene expression throughout this developmental program, the Δ*CD2758* mutant may appear to initiate sporulation less efficiently than the wild-type because it is down-regulating sporulation at the 19 h time point, leaving expression of only late sporulation genes such as SigK-dependent genes. This model would reconcile the decreased spore titers and sporulation efficiencies observed (Fig. 2, Supplementary Fig. 2b) with the transcriptional data, and is consistent with previous observations made in *Bacillus subtilis* sporulation[37].

Measuring sporulation gene expression at earlier time points would presumably better reflect the functional differences measured between Δ*CD2758* relative to wild-type.

We observed a significant enrichment of CAAAAA motif sites within DE genes compared to non-DE ones ($P < 10^{-3}$, Mann-Whitney-Wilcoxon test) (Supplementary Fig. 6c) independently of the time point considered (Supplementary Fig. 6d). More specifically, we observed a significant association between the number of CAAAAA target sites within the coding and regulatory regions of DE genes and their expression change in Δ*CD2758* cells, especially for genes containing more than 4 target sites ($P < 0.05$, Mann-Whitney-Wilcoxon test) (Supplementary Figs. 6e-h).

Hence, these results consolidate the involvement of CD2758 in the sporulation process (particularly visible at the 19 h sporulation time point) and provide molecular evidence supporting reduced sporulation initiation in Δ*CD2758* cells.

Integrative transcriptomic analyses suggest epigenetic regulation of *C. difficile* colonization and biofilm formation

Considering the high conservation of CD2758 across *C. difficile* genomes, we attempted to explore the RNA-seq data further beyond the sporulation phenotype and the GO term analyses discussed above, with a special focus on biological processes critical to *C. difficile* infection. At the single gene level, we examined the two *C. difficile* toxins (A and B) in the RNA-seq data and found that *tcdA* has significantly increased expression in the exponential (fold change = 1.6; $P < 10^{-10}$) and stationary phases (fold change = 1.7; $P < 10^{-12}$) in the mutant relative to wild-type, and *tcdB* has significantly decreased expression 15 h post-sporulation induction (fold change = 0.73; $P < 10^{-3}$), consistent with qPCR-based validation (Supplementary Fig. 7a). At the gene set level, we took an integrative transcriptomic strategy to search for critical *C. difficile* biological processes that may involve epigenetic regulation.

Specifically, we performed an overlap analysis between the list of DE genes from our RNA-seq data (wild-type vs. Δ*CD2758* mutant; four different time points), and those from published studies focusing on the colonization and infection by this pathogen (Materials and Methods). Out of the total 20 pairwise overlap analyses, we observed 7 significant overlaps ($P < 0.01$ with Bonferroni correction) between our dataset and those of others (Supplementary Table 7e).

First, using DE genes obtained from murine gut isolates at increasing time points after infection[38], we found significant overlaps with DE genes in the Δ*CD2758* mutant (O/E=1.8, $P < 10^{-4}$, Chi-Square test) (Fig. 5b, Supplementary Table 7e). Second, by comparing with genes DE at different murine gut microbiome compositions[39], we found a statistically significant overlap, particularly involving genes expressed at stationary stage (O/E=1.4, $P < 10^{-6}$, Chi-Square test) (Supplementary Fig. 7b, Supplementary Table 7e). Finally, DE genes in the Δ*CD2758* mutant (sporulation phase) had a significant overlap compared to DE genes in conditions favoring the production of biofilms[40] (Supplementary Fig. 7c, Supplementary Table 7e) (O/E=1.3, $P < 10^{-9}$, Chi-Square test). Collectively, these integrative overlap analyses provide additional evidence that DNA methylation events by CD2758 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.

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

Until now, we mainly concentrated our attention on the CD2758 MTase and on its methylation motif CAAAAA. In this section we extended our analysis to the multiple R-M systems associated with the *C. difficile* epigenome. As a starting point, we made use of i) the unique information on R-M recognition motifs provided by SMRT-seq, and ii) the exceptional diversity of Type I R-M systems (as well as the near depletion of other types of complete

systems) observed in our *C. difficile* dataset, to better understand their impact on phage target side avoidance and gene flux. Restriction site avoidance is the most effective way to escape the action of R-M systems, and it has been predominantly studied for those belonging to Type II systems[41-44]. To investigate this, we selected representative members of the *Siphoviridae* and *Myoviridae* families and used Markov chain models to compute the number of observed and expected Type I R-M target sites accounting for oligonucleotide composition of each phage genome (Materials and Methods). We then used as a measure of genetic transfer, the number of recent horizontal gene transfer (HGT) gains from the pattern of presence/absence of gene families in the species tree[45]. Our data suggest that *C. difficile* phages have generally evolved to reduce the number of several Type I recognition sites (Fig. 6a). Concomitantly, we found an inverse linear trend between HGT and O/E ratios of Type I motif targets in phages (Spearman's ρ = -0.826, $P < 0.05$) (Fig. 6b). This suggests a link between the frequencies at which different *C. difficile* genomes (or STs) are targeted by phages, and the latter's capacity to underrepresent certain motifs targeted by the cell's R-M machinery.

While the relationship between some defense systems (e.g. CRISPR-Cas and R-M systems) has been studied in an experimental setting[46,47], the rich collection of *C. difficile* methylomes and their genomes provides an unprecedented opportunity to jointly analyze its diversity of defense systems in a data-driven manner. In addition to R-M systems, we searched for evidence of additional systems: CRISPR-Cas, toxin-antitoxin (T-A), abortive infection (Abi) systems, bacteriophage exclusion (BREX)[48], prokaryotic Argonautes (pAgos)[49], DISARM[50], and a set of 10 recently-discovered defense systems[51] (Materials and Methods). Only T-A and CRISPR-Cas systems were ubiquitous in our genomes (Fig. 6c, Supplementary Tables 8a-d, Supplementary Figs. 8a, c, d), and all CRISPR-Cas systems detected were of Type-IB [52], consistent with earlier studies[53-55] (Supplementary Fig. 8b).

Different types of defense systems are expected to confer different degrees of protection against invading DNA. Also, under certain conditions, some defense systems may even facilitate genetic exchange between cells, as recently shown for CRISPR-Cas[56] and R-M systems[12]. Thus, we enquired how gene flux was distributed within our dataset (Materials and Methods; Fig. 6c, Supplementary Table 8e, f; Supplementary Text), and how is it associated with the multiple defense systems present in it. Specifically, to test the effect of R-Ms, CRISPR-Cas, and T-As (the most abundant defense systems found across *C. difficile* strains) on gene flux, we built stepwise linear models to assess the role of each of these variables in explaining the variance of HGT and HR (Supplementary Table 8g). Interestingly, we found a strong positive association between CRISPR spacer count and HGT/HR (Supplementary Table 8g, Supplementary Figs. 9a, b). The increase in gene flux with spacer content, although somehow unexpected, could be reconciled if generalized transduction occurs in *C. difficile*[57,58], as recently shown for *Pectobacterium*[56]. R-M and T-A abundance had a less important explanatory role in the prevention of gene flux (Supplementary Table 8g).

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[8]. Because CDI 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 the first comprehensive characterization of the DNA methylation landscape across a diverse collection of clinical isolates. During our epigenome analysis, we identified an m6A MTase (*CD2758*) 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 both *in vitro* and *in vivo* (Fig. 2).

Multiple factors likely contribute to the more pronounced effect in the mouse model: 1) the MTase may play a more important role in *C. difficile* sporulation *in vivo*, 2) in the mouse model, the post-infection spore measurement reflects an accumulate effect, 3) the MTase likely has pleiotropic effects beyond sporulation, consistent with our integrative RNA-seq data analysis, and these additional effects also contribute, indirectly, to the reduced sporulation. This represents, to the best of our knowledge, the first time that DNA methylation has been found to impact sporulation in any bacterium, opening a new dimension to study *C. difficile*.

The highly conserved *CD2758* and its flanking genes across *C. difficile* genomes suggest that additional phenotypes may be regulated by CD2758 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 CD2758, was the significant overlap between transcriptional signatures between our study (WT vs. *CD2758* mutant) and those of others observed during the *in vivo* colonization and biofilm formation (Fig. 5b, Supplementary Figs. 7b, c).

The fact that *CD2758* is a solitary MTase 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 m6A MTase was Dam targeting GATC in *E. coli*. Dam plays multiple important functions and is essential in some pathogens[31]. However, since it is conserved in the large diversity of γ-proteobacteria, it was not considered a promising drug target. In contrast, the uniqueness of *CD2758* 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. Since this MTase seemingly does not impact the general fitness of *C. difficile*[25], a drug specifically targeting it may be developed with a lower chance for resistance.

Considering the large number of genes differentially expressed in the CD2578 mutant, the functional impact of CAAAAA methylation is likely mediated by multiple genes, either directly regulated by DNA methylation or indirectly through the transcriptional cascade. Mechanistically, DNA methylation can either activate or repress a gene depending on other DNA binding proteins that compete with DNA MTases[14,31] Specifically, the competition between transcription factors and MTases may form an epigenetic switch to turn on/off a gene. In our analysis, some non-methylated CAAAAA sites are conserved across multiple genomes. One such site corresponds to the promoter region of *tcdR* (target of transcription factor CodY[59]), a gene that codes for an alternate sigma factor that initiates transcription of the *tcdA* and *tcdB* toxin genes of the PaLoc region. Previous evidence showed that CodY represses toxin gene expression by binding with high affinity to the *tcdR* promoter region[59]. Building on this, our results suggest that the regulatory networks of CD2758 and those of toxin production may be linked together in *C. difficile*. These observations are interesting clues that can be pursued in future work, although the complexity of epigenetic regulation may be beyond our current understanding. For example, our analysis showed that genes containing 5 or more motifs showed the greatest increase in expression in the Δ*CD2758* mutant, suggesting that the effects of methylation may be additive within a single gene and/or that a threshold level of methylation may be present for regulation to occur.

Our analysis shows that the *C. difficile* genome is characterized by a large and diverse pan-genome and low levels of genome conservation, which fall in the interval of previous estimates[60-62]. Both HR and HGT play a significant role in sequence diversification (including at CAAAAA sites), confirming the dynamic nature of *C. difficile's* genome. Despite the multitude of defense systems present in this bacterium, gene flux was mainly correlated with CRISPR spacer abundance and, to a lesser extent, with R-M abundance. Methylation motif information obtained by SMRT-seq also allowed computing motif avoidance in known *C. difficile* phages. We found that *C. difficile* genomes undergoing higher HGT and HR are

typically targeted by phages avoiding methylation motifs targeted by the host's Type I R-M machinery.

With more than 1,500 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 the first comprehensive comparative analysis of a large collection of a single bacterial species and thus provides a detailed roadmap that can be used by the scientific community to leverage the current status quo of epigenetic analyses.

Materials and Methods

Data and code availability

Genome assemblies will be available via NCBI under BioProject ID PRJNA448390. RNA-Seq data has been submitted to the NCBI Sequence Read Archive (SRA) under project PRJNA445308. Scripts and a tutorial supporting all key analyses of this work will be made publically available at http://github.com/fanglab/ upon the acceptance of the manuscript.

Clostridium difficile isolates and culture

36 clonal *C. difficile* isolates from CDI 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[10] (Supplementary Table 3).

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 0.1ºC 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 hours 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 minutes 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. 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.

Identification of defense systems

Identification of R-M systems was performed as previously described[11]. 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[63] (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[64]. Each protein family was aligned with MAFFT v7.305b[65] using the E-INS-i option, 1,000 cycles of iterative refinement, and offset 0. Alignments were visualized in SEAVIEW v4.6.1[66] and manually trimmed to remove poorly aligned regions at the extremities. Hidden Markov model (HMM) profiles were then built from each multiple sequence alignment using the hmmbuild program from the HMMER v3.0 suite[67] (default parameters) (available at https://github.com/pedrocas81). 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[68] 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[69] 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[48]. DISARM systems were identified using the PFAM signature domains (PF09369, PF00271, PF13091) belonging to the core gene triplet characteristic of this system[50]. To search for prokaryotic Argonaute (pAgo) genes we built a dedicated HMM profile based on a list of 90 Ago-PIWI proteins[70]. Searches for the ensemble of newly found antiphage systems were performed using the list of PFAM profiles published by the authors[51]. Type II toxin-antitoxin (T-A) systems were detected using the TAFinder tool[71] with default parameters.

CAAAAA motif abundance and exceptionality

We evaluated the exceptionality of the CAAAAA motif using R'MES[29] v3.1.0 (http://migale.jouy.inra.fr/?q=rmes). 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[30] 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[72] v2.4.0 with default parameters. Since progressiveMauve does not rely on annotations to guide the alignment, we first used the Mauve Contig Mover[73] 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[74] 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 (http://regprecise.lbl.gov[75], last accessed July 2017). These were converted to PWMs using in-house developed scripts. This led to a total of 21 PWMs pertaining to 14 distinct transcription factor families (Supplementary Table 6b). Matches between these matrices and *C. difficile* genomes was performed with MAST[76] (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 hits those with *P*-values $<10^{-8}$. Transcription start sites (TSSs) were predicted with Parseq[33] under the 'fast' speed option from eight 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 sequencing, read alignment, and differential expression analysis

Purified mRNA was extracted from three biological replicates of exponential, stationary, and sporulating (15 and 19 h) *C. difficile* 630Δ*erm* and *C. difficile* 630Δ*erm*Δ*CD2758*, and converted to cDNA as previously described[77]. 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. Read quality was checked using FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) (no reads flagged as having poor quality or as containing adapter sequences were used in downstream analyses). Reads were aligned to the *C. difficile* 630 reference genome using BWA[78]. 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[79]. Genes with a *P*-value $<10^{-2}$, and a false discovery rate (FDR) $<10^{-2}$ were called as differentially expressed. Functional classification of genes was performed using the DAVID online database (https://david.ncifcrf.gov)[80] with Fisher's exact test enrichment statistics, a

Benjamini-Hochberg corrected *P*-value cutoff of 0.05, and a FDR < 0.05. 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[38-40], 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* [81].

Identification of core- and pan-genome

The *C. difficile* core-genome was built using a methodology previously published[82]. 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). These parameters (four genes being less than one-half of the diameter of the neighborhood) allow retrieving orthologs on the edge of rearrangement breakpoints and therefore render the analysis robust to the presence of rearrangements. 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 (v1.2.11, http://lbbe.univ-lyon1.fr/SiLiX)[83]. 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[84] 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(α = 1 - δ) 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 v10.7.5[85] and Geneconv v1.81a[86]. The first used a predefined tree (*i.e*. the clade'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[85]. Mean patristic branch lengths were computed with the R package "ape" v3.3[87], and transition/transversion ratios were computed with the R package "PopGenome" v2.1.6[88]. 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[45] (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 HTG 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 $\rho = 0.96$, $P < 10^{-4}$).

References

1. Smits, W. K., Lyras, D., Lacy, D. B., Wilcox, M. H. & Kuijper, E. J. *Clostridium difficile* infection. *Nat. Rev. Dis. Primers* **2**, 16020 (2016).

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

3. Seekatz, A. M. & Young, V. B. *Clostridium difficile* and the microbiota. *J. Clin. Invest.* **124**, 4182-4189 (2014).

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

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

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

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

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

9. Sebaihia, M.*, et al.* The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat. Genet.* **38**, 779-786 (2006).

10. He, M.*, et al.* Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat. Genet.* **45**, 109-113 (2013).

11. Oliveira, P. H., Touchon, M. & Rocha, E. P. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **42**, 10618-10631 (2014).

12. Oliveira, P. H., Touchon, M. & Rocha, E. P. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl. Acad. Sci. USA* **113**, 5658-5663 (2016).

13. Casadesus, J. & Low, D. Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.* **70**, 830-856 (2006).

14. Low, D. A., Weyand, N. J. & Mahan, M. J. Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. *Infect. Immun.* **69**, 7197-7204 (2001).

15. Cohen, N. R.*, et al.* A role for the bacterial GATC methylome in antibiotic stress survival. *Nat. Genet.* **48**, 581-586 (2016).

16. Manso, A. S.*, et al.* A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nat. Commun.* **5**, 5055 (2014).

17. Atack, J. M.*, et al.* A biphasic epigenetic switch controls immunoevasion, virulence and niche adaptation in non-typeable *Haemophilus influenzae*. *Nat. Commun.* **6**, 7828 (2015).

18. Flusberg, B. A.*, et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461-465 (2010).

19. 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).

20. Murray, I. A.*, et al.* The methylomes of six bacteria. *Nucleic Acids Res.* **40**, 11450-11462 (2012).

21. Davis, B. M., Chao, M. C. & Waldor, M. K. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. *Curr. Opin. Microbiol.* **16**, 192-198 (2013).

22. Smits, W. K. Hype or hypervirulence: a reflection on problematic *C. difficile* strains. *Virulence* **4**, 592-596 (2013).

23. Roberts, R. J.*, et al.* A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* **31**, 1805-1812 (2003).

24. Wust, J., Sullivan, N. M., Hardegger, U. & Wilkins, T. D. Investigation of an outbreak of antibiotic-associated colitis by various typing methods. *J. Clin. Microbiol.* **16**, 1096-1101 (1982).

25. Dembek, M.*, et al.* High-throughput analysis of gene essentiality and sporulation in *Clostridium difficile*. *MBio* **6**, e02383 (2015).

26. Ng, Y. K.*, et al.* Expanding the repertoire of gene tools for precise manipulation of the *Clostridium difficile* genome: allelic exchange using *pyrE* alleles. *PLoS ONE* **8**, e56051 (2013).

27. Donnelly, M. L., Fimlaid, K. A. & Shen, A. Characterization of *Clostridium difficile* spores lacking either SpoVAC or dipicolinic acid synthetase. *J. Bacteriol.* **198**, 1694-1707 (2016).

28. Chen, X.*, et al.* A mouse model of *Clostridium difficile*-associated disease. *Gastroenterology* **135**, 1984-1992 (2008).

29. 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* (eds L. Elnitsk, H. Piontkivska, and L. Welch). World Scientific (2011).

30. Knijnenburg, T. A.*, et al.* Multiscale representation of genomic signals. *Nat. Methods* **11**, 689-694 (2014).

31. Wion, D. & Casadesus, J. N6-methyl-adenine: an epigenetic signal for DNA-protein interactions. *Nat. Rev. Microbiol.* **4**, 183-192 (2006).

32. Lim, H. N. & van Oudenaarden, A. A multistep epigenetic switch enables the stable inheritance of DNA methylation states. *Nat. Genet.* **39**, 269-275 (2007).

33. 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).

34. Fimlaid, K. A.*, et al.* Global analysis of the sporulation pathway of *Clostridium difficile*. *PLoS Genet.* **9**, e1003660 (2013).

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).

36. Fimlaid, K. A. & Shen, A. Diverse mechanisms regulate sporulation sigma factor activity in the *Firmicutes*. *Curr. Opin. Microbiol.* **24**, 88-95 (2015).

37. Tan, I. S. & Ramamurthi, K. S. Spore formation in *Bacillus subtilis*. *Environ. Microbiol. Rep.* **6**, 212-225 (2014).

38. Fletcher, J. R., Erwin, S., Lanzas, C. & Theriot, C. M. Shifts in the gut metabolome and *Clostridium difficile* transcriptome throughout colonization and infection in a mouse model. *mSphere* **3**, (2018).

39. Jenior, M. L., Leslie, J. L., Young, V. B. & Schloss, P. D. *Clostridium difficile* colonizes alternative nutrient niches during infection across distinct murine gut microbiomes. *mSystems* **2**, (2017).

40. Maldarelli, G. A.*, et al.* Type IV pili promote early biofilm formation by *Clostridium difficile*. *Pathog. Dis.* **74**, (2016).

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

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

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

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

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

46. Yaung, S. J., Esvelt, K. M. & Church, G. M. CRISPR/Cas9-mediated phage resistance is not impeded by the DNA modifications of phage T4. *PLoS ONE* **9**, e98811 (2014).

47. Dupuis, M. E., Villion, M., Magadan, A. H. & Moineau, S. CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. *Nat. Commun.* **4**, 2087 (2013).

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

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

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

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

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

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

54. Hargreaves, K. R., Flores, C. O., Lawley, T. D. & Clokie, M. R. Abundant and diverse clustered regularly interspaced short palindromic repeat spacers in *Clostridium difficile* strains and prophages target multiple phage types within this pathogen. *MBio* **5**, e01045-01013 (2014).

55. Andersen, J. M., Shoup, M., Robinson, C., Britton, R., Olsen, K. E. & Barrangou, R. CRISPR diversity and microevolution in *Clostridium difficile*. *Genome Biol. Evol.* **8**, 2841-2855 (2016).

56. Watson, B. N. J., Staals, R. H. J. & Fineran, P. C. CRISPR-Cas-mediated phage resistance enhances horizontal gene transfer by transduction. *MBio* **9**, (2018).

57. Hargreaves, K. R. & Clokie, M. R. *Clostridium difficile* phages: still difficult? *Front. Microbiol.* **5**, 184 (2014).

58. Goh, S., Hussain, H., Chang, B. J., Emmett, W., Riley, T. V. & Mullany, P. Phage varphiC2 mediates transduction of Tn6215, encoding erythromycin resistance, between *Clostridium difficile* strains. *MBio* **4**, e00840-00813 (2013).

59. Dineen, S. S., McBride, S. M. & Sonenshein, A. L. Integration of metabolism and virulence by *Clostridium difficile* CodY. *J. Bacteriol.* **192**, 5350-5362 (2010).

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

61. Scaria, J., Ponnala, L., Janvilisri, T., Yan, W., Mueller, L. A. & Chang, Y. F. Analysis of ultra low genome conservation in *Clostridium difficile*. *PLoS ONE* **5**, e15147 (2010).

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

63. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **43**, D298-299 (2015).

64. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575-1584 (2002).

65. Katoh, K. & Standley, D. M. MAFFT: iterative refinement and additional methods. *Methods Mol. Biol.* **1079**, 131-146 (2014).

66. 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).

67. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29-37 (2011).

68. Bland, C.*, et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinf.* **8**, 209 (2007).

69. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371-373 (2003).

70. Makarova, K. S., Wolf, Y. I., van der Oost, J. & Koonin, E. V. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biol. Direct* **4**, 29 (2009).

71. Xie, Y.*, et al.* TADB 2.0: an updated database of bacterial type II toxin-antitoxin loci. *Nucleic Acids Res.* **46**, D749-D753 (2018).

72. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).

73. Rissman, A. I., Mau, B., Biehl, B. S., Darling, A. E., Glasner, J. D. & Perna, N. T. Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* **25**, 2071-2073 (2009).

74. Danecek, P.*, et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

75. Novichkov, P. S.*, et al.* RegPrecise 3.0--a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics* **14**, 745 (2013).

76. Bailey, T. L. & Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48-54 (1998).

77. Chao, M. C.*, et al.* Correction: A cytosine methytransferase modulates the cell envelope stress response in the *Cholera* pathogen. *PLoS Genet.* **11**, e1005739 (2015).

78. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

79. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

80. Huang, D. W.*, et al.* DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* **35**, W169-175 (2007).

81. Wang, M., Zhao, Y. & Zhang, B. Efficient test and visualization of multi-set intersections. *Sci Rep* **5**, 16923 (2015).

82. Touchon, M.*, et al.* Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* **5**, e1000344 (2009).

83. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* **12**, 116 (2011).

84. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **11**, 472-477 (2008).

85. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).

86. Sawyer, S. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**, 526-538 (1989).

87. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289-290 (2004).

88. Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**, 1929-1936 (2014).

89. Bordeleau, E., Fortier, L. C., Malouin, F. & Burrus, V. c-di-GMP turn-over in *Clostridium difficile* is controlled by a plethora of diguanylate cyclases and phosphodiesterases. *PLoS Genet.* **7**, e1002039 (2011).

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

Acknowledgements

We acknowledge Eduardo P.C. Rocha (Institut Pasteur, Paris, France) for critical reading and for providing helpful comments/suggestions. 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. 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.

Author Contributions

G.F. conceived and supervised the project. P.H.O. and G.F. designed the methods. P.H.O. performed most of the computational analyses and developed most of the scripts supporting the analyses. A.Kim and G.F. developed the motif refinement tool. P.H.O., A.S., R.T. and G.F. contributed to experimental design. A.S. constructed the deletion and catalytic *CD2758* mutants, performed complementation, sporulation and phenotypic assays. O.S. and E.A.M. performed qRT-PCR controls for RNA-seq analyses. E.M.G., D.T. and R.T. performed the mouse infection experiment and analyzed the data. O.S., E.A.M., G.D., M.L., C.B., N.Z., D.A., I.O., G.P., 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., A.Kim., O.S., T.P., S.Z., E.A.M., M.T., A.A., A.B., R.J.B., R.T., E.E.S., R.S., H.B., A.Kasarskis., A.S. and G.F. analyzed the data. P.H.O. and G.F. wrote the manuscript with writing contributions provided by A.S. and R.T and additional information inputs from all co-authors.

Figure Legends

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 (inside the chromosome). No hits were obtained in plasmids. (f) Immediate genomic context of *CD2758*. 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 non-catalytically active GGDEF domain (PF00990) and a cache domain (PF02743); *ptsI*, *ptsH* belong to a phosphotransferase (PTS) system; *CD2757*: patatin-like phospholipase (PF01734); *CD2758*: Type II MTase; *CD2759*: Rrf2-type transcriptional regulator; *CD2760*: phosphodiesterase with a non-catalytically active GGDEF domain (PF00990) and a conserved EAL domain (PF00563)[89]; *CD2761*: N-acetylmuramoyl-L-alanine amidase; *CD2762*: undecaprenyl diphosphate synthase.

Fig. 2. CD2758 modulates sporulation levels in *C. difficile*. (a) Spore purification efficiencies obtained from sporulating cells harvested 3 days after plating. 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 water in which spores were re-suspended. (b) Heat-resistance (H_{RES}) efficiencies of sporulating cultures 22 h after sporulation induction were determined relative to wild-type. Statistical analyses were performed with a one-way ANOVA with Tukey's test. (c) 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[90]. 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. (d) Kinetics of infection in mice (*n*=6) following inoculation with a sub-lethal amount (10^5 spores) of wild type (WT) *C. difficile* 630*Δerm*, MTase mutant *ΔermΔCD2758*, and complement *ΔermΔCD2758-C*. Empty symbols indicate that all animals in that group yielded CFUs below the limit of detection (indicated by the dotted line). Log_{10}-transformed data from each time point were analyzed by ANOVA. ** $P < 10^{-2}$, *** $P < 10^{-3}$, **** $P < 10^{-4}$.

Fig. 3. Abundance, distribution, and conservation of CAAAAA motif sites. (a) Distribution of CAAAAA motif sites in both strands of the reference *C. difficile* 630 genome, and corresponding genomic signal obtained by MSR. Letters (A-E) represent regions with particularly high abundance of CAAAAA motifs at scales above 20, i.e., typically above the single gene level (median gene size in *C. difficile* = 0.78 kb) (see Supplementary Table 5d for genes contained in each of the regions). 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 $P = 10^{-6}$. (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 occurrences of CAAAAA are indicated as orange-shaded regions. The results are shown in Fig. 3a in the form of heatmaps. (c) DAVID functional analysis of the genes containing intragenic and regulatory (100 bp upstream the start codon) orthologous variable CAAAAA motif sites. We have considered a Fisher's exact test enrichment statistics, a Benjamini-Hochberg corrected *P*-value cutoff of 0.05, and a false discovery rate (FDR) < 0.05.

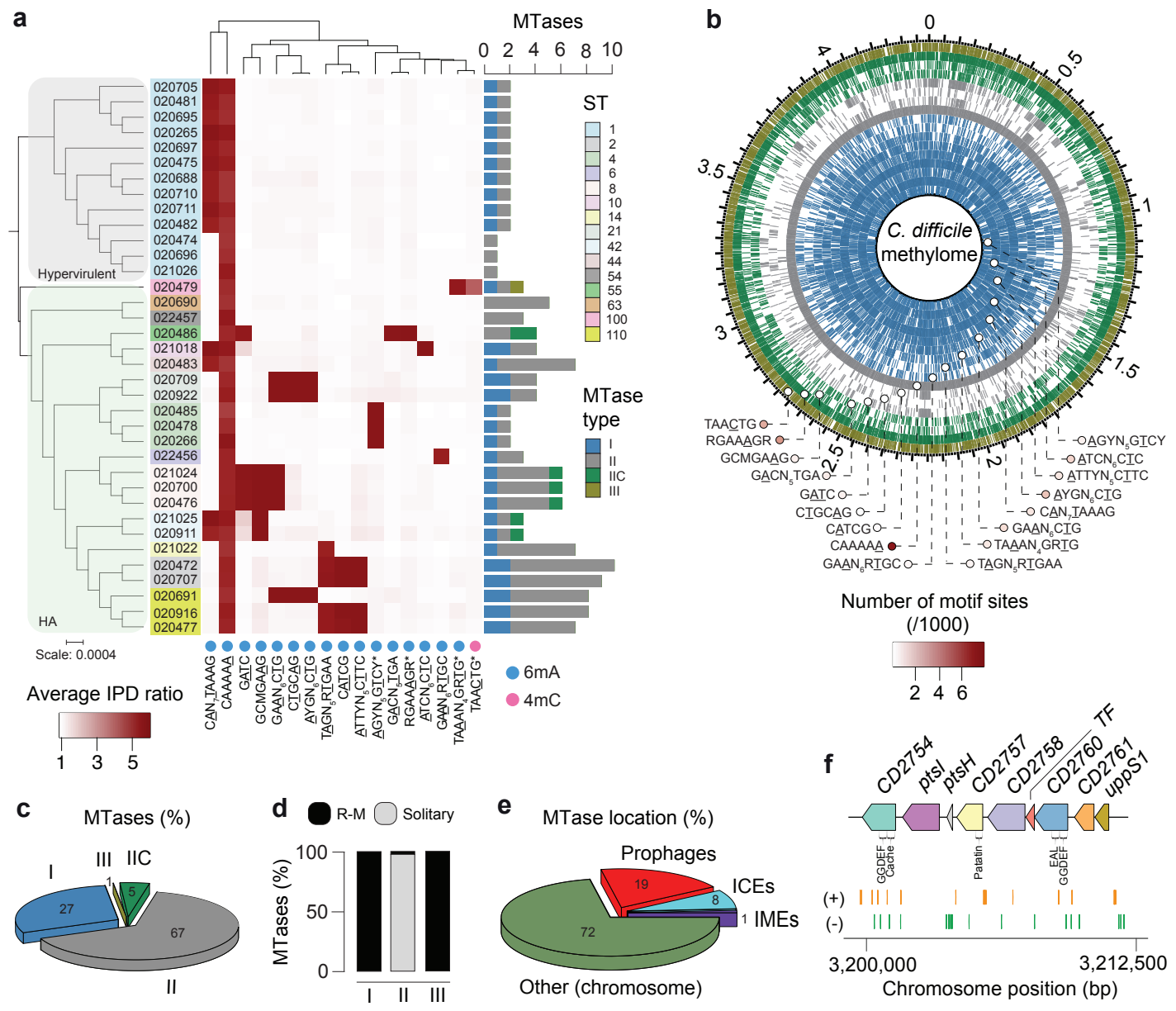
Fig. 4. Distribution of non-methylated CAAAAA motif sites, and overlap with transcription factor binding sites (TFBS) and transcription start sites (TSS). (a) Number of *C. difficile* isolates for which non-methylated CAAAAA motif sites were detected at a given chromosome position (coordinates are relative 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, blue forms). Other examples of conserved non-methylated CAAAAA motif sites are illustrated in Supplementary Fig. 4b. (c) % CAAAAA motif sites (non-methylated and methylated) overlapping CodY and XylR for each *C. difficile* isolate. (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 *C. difficile* isolate. (f) % mutated reads (SNPs + indels) in CAAAAA and K(G or T)AAAAA motifs for *M. tuberculosis* (MT), *E. coli* (EC), *C. difficile* (CD) and *H. pylori* (HP). AAAAAA was not considered as control motif as it would theoretically be more error-prone. *** $P < 10^{-3}$, Mann-Whitney-Wilcoxon test.

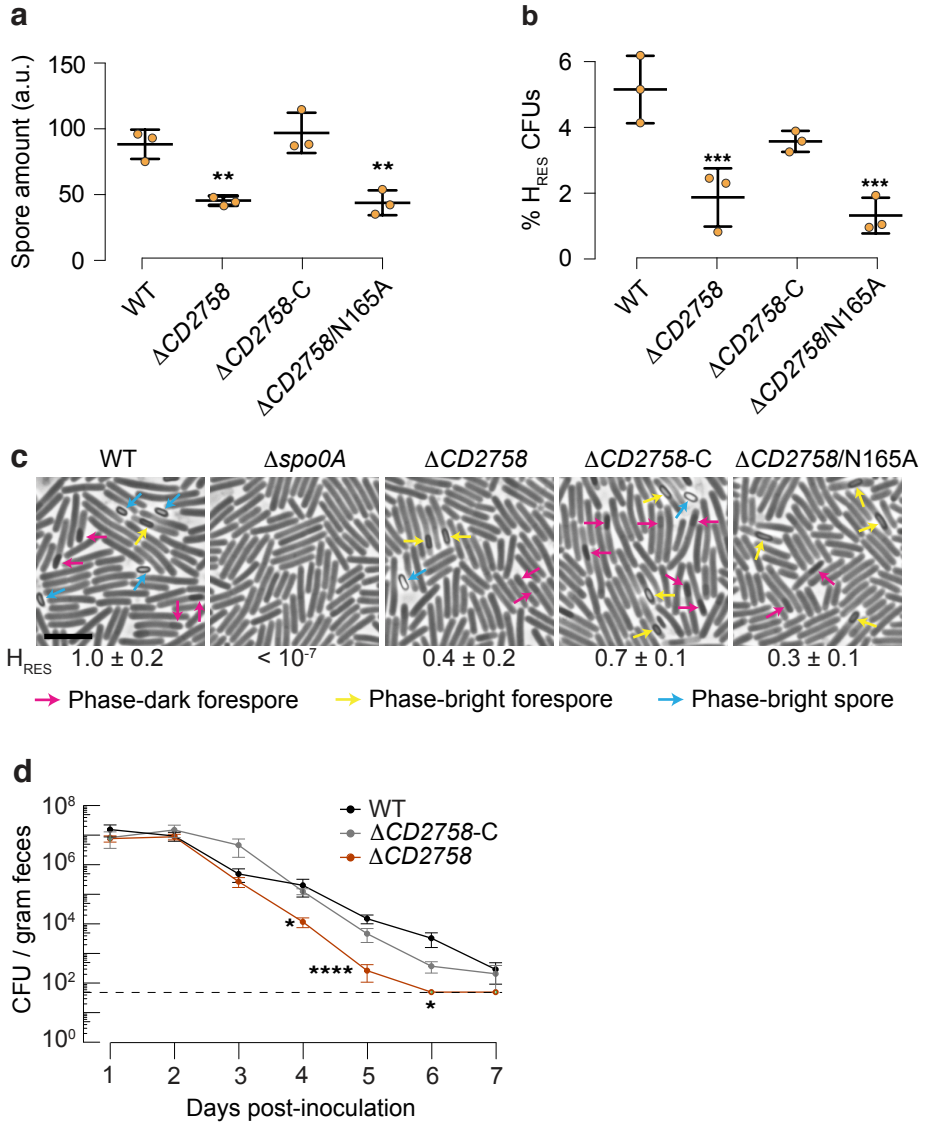
Fig. 5. Gene expression analysis. (a) Heatmap of 92 DE genes in three replicates of *C. difficile* 630Δ*erm* compared to equal number of replicates of *C. difficile* 630Δ*erm*Δ*CD2758* and that are in the pathways indicated in Supplementary Table 7c. The *Z* score reflects the

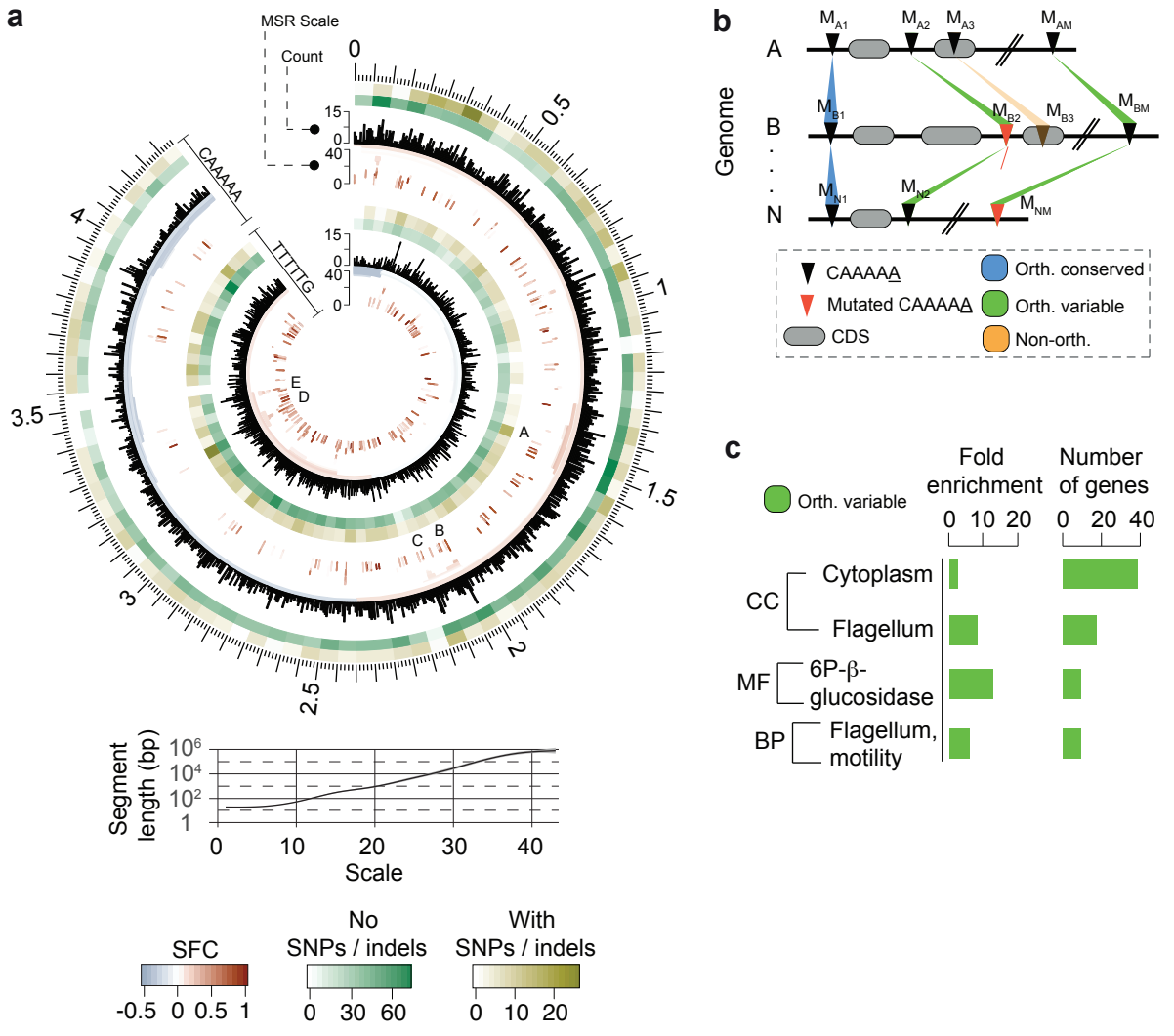
degree of down- (*Z* score < 0) or up- (*Z* score > 0) regulation, computed by subtracting the mean of the log-transformed expression values and dividing by the standard deviation for each gene over all samples scored. (b) Significance of overlap between multiple datasets of DE genes. Comparisons were performed between DE genes called in this study for each time point (blue-shaded) and those from Fletcher *et al*.[38] (green-shaded). The latter corresponds to *C. difficile* DE genes called 24 and 36 h (versus 12 h) of post-infection time in a murine model. 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). Stars indicate pairwise comparisons between the different studies.

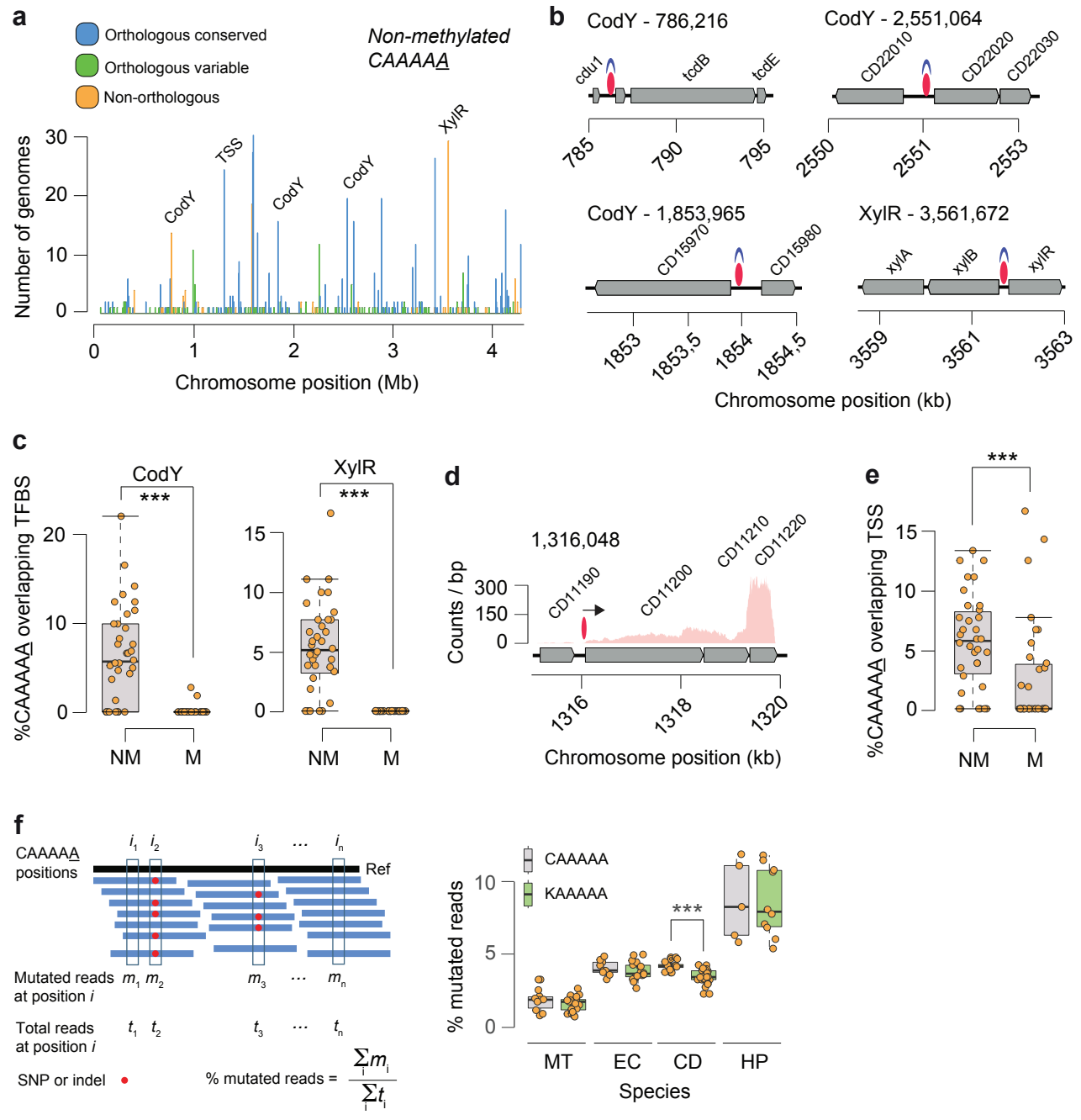
Fig. 6. Interplay between defense systems and gene flux in *C. difficile*. (a) Observed/expected (O/E) ratios for Type I target recognition motifs in *Clostridium* phage genomes. 6 phage genomes representative of *Siphoviridae* and *Myoviridae* families and tail types were analyzed (phiCD111, phiCDHM11, phiMMP01, phiMMP04, phiC2, phiCD38). For each motif, we tested if the median value of the O/E ratio in phage genomes was significantly different from 1 with the Mann-Whitney test. O/E values were obtained with R'MES using Markov chain models that take into consideration oligonucleotide composition (Materials and Methods). (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 motifs indicated in brackets. (c) Heatmap aggregate depicts: abundance of defense systems (R-M, average number of spacers per CRISPR, T-A, Abi, Shedu systems), 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. Searches

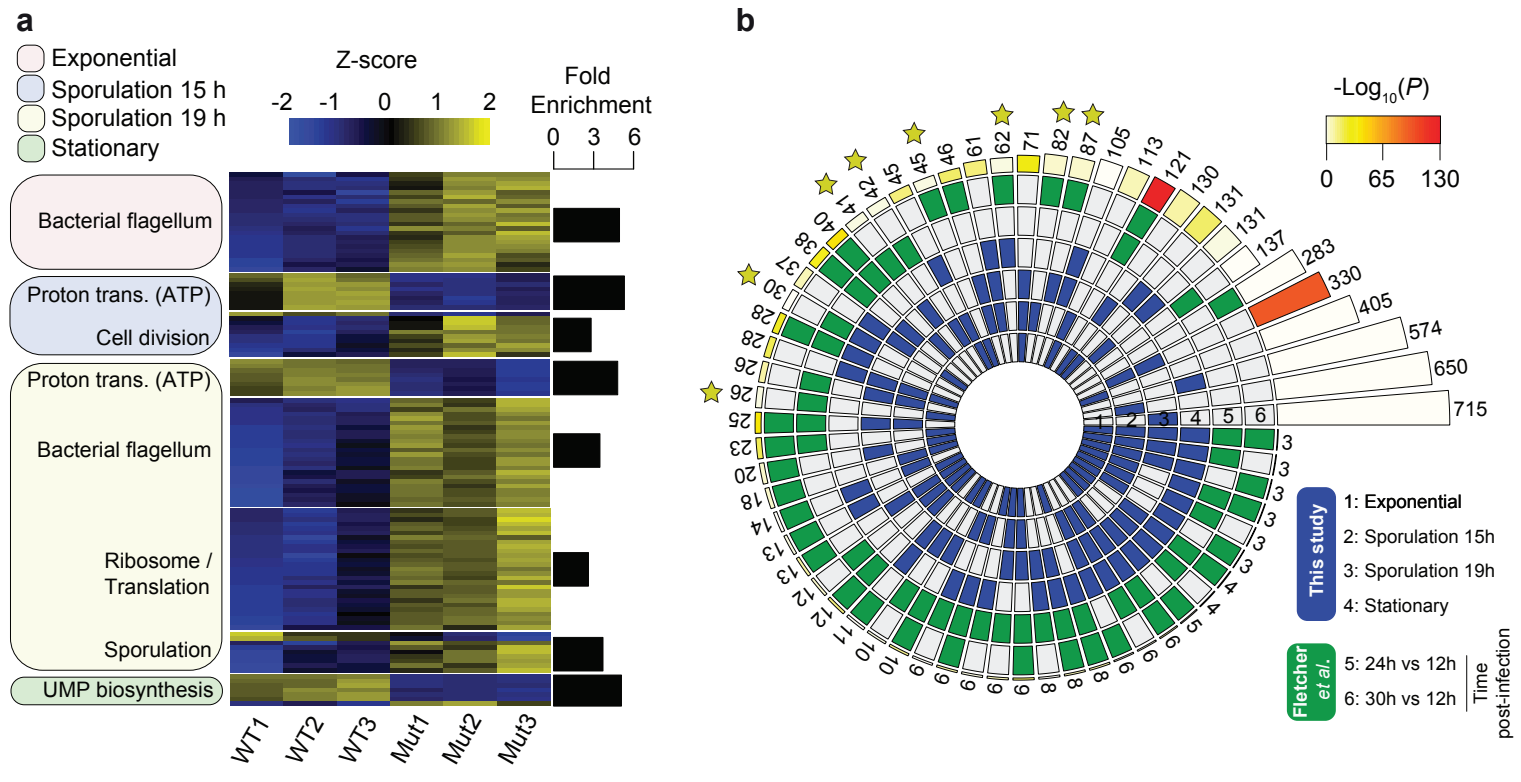
were performed against well-known *C. difficile* phages comprising: five siphophages (ϕCD111 / NC_028905.1, CD146 / NC_028958.1, CD38-2 / NC_015568.1, CD6356, CD211), five small-tail myophages (MMP04, CD506, CDHM11, CD481-1, CDHM13), five medium-tail myophages (MMP03, CDMH1, C2, CD119, CDHM19), and four long-tail myophages (CD2, MMP02, CD50, MMP01). Phages were clustered according to their family (*Siphoviridae* (S), *Myoviridae* (M)), and tail type. * $P<0.05$, ** $P<10^{-2}$, *** $P<10^{-3}$, Mann-Whitney-Wilcoxon test.

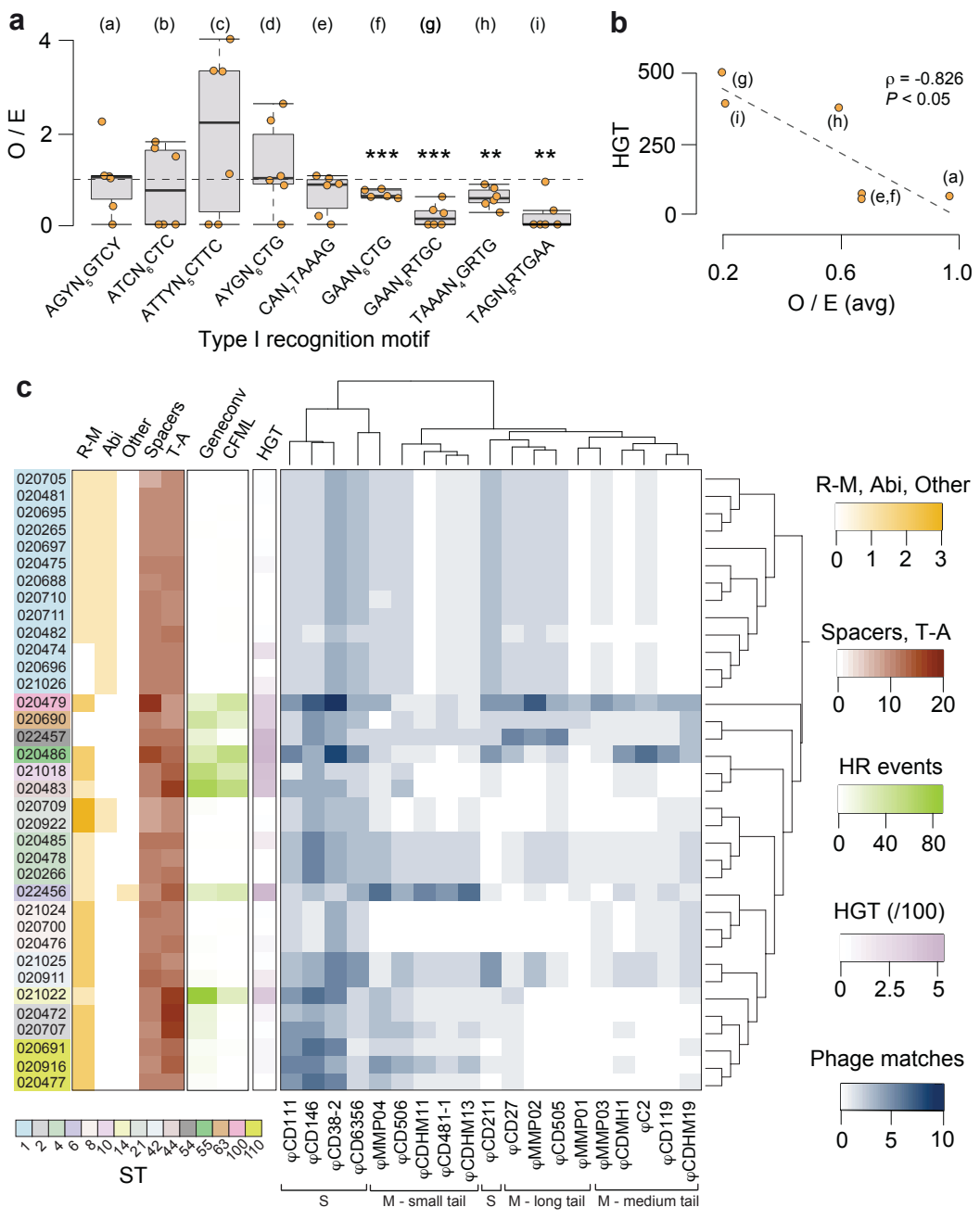

a
MTases
0 2 4 6 8 10
ST
1
2
4
6
8
10
14
21
42
44
54
55
63
100
110
Hypervirulent
HA
020705
020481
020695
020265
020697
020475
020688
020710
020711
020482
020474
020696
021026
020479
020690
022457
020486
021018
020483
020709
020922
020485
020478
020266
022456
021024
020700
020476
021025
020911
021022
020472
020707
020691
020916
020477
Scale: 0.0004
MTase type
I
II
IIC
III
CAN7TAAAG
CAAAAA
GATC
GCMGAAG
GAAN6CTG
CTGCAG
AYGN6CTG
TAGN5RTGAA
CATCG
ATTYN5CTTC
AGYN5GTCY*
GACN5TGA
RGAAAGR*
ATCN6CTC
GAAN6RTGC
TAAAN4GRTG*
TAACTG*
6mA
4mC
Average IPD ratio
1 3 5
b
0
0.5
1
1.5
2
2.5
3
3.5
4
C. difficile
methylome
TAACTG
RGAAAGR
GCMGAAG
GACN5TGA
GATC
CTGCAG
CATCG
CAAAAA
GAAN6RTGC
AGYN5GTCY
ATCN6CTC
ATTYN5CTTC
AYGN6CTG
CAN7TAAAG
GAAN6CTG
TAAAN4GRTG
TAGN5RTGAA
Number of motif sites
(/1000)
2 4 6
c
MTases (%)
I
III
IIC
II
27
1
5
67
d
R-M
Solitary
MTases (%)
100
50
0
I
II
III
e
MTase location (%)
Prophages
ICEs
IMEs
Other (chromosome)
19
8
1
72
f
CD2754
ptsI
ptsH
CD2757
CD2758
CD2760
CD2761
uppS1
TF
GGDEF
Cache
Patatin
EAL
GGDEF
(+)
(-)
3,200,000
3,212,500
Chromosome position (bp)

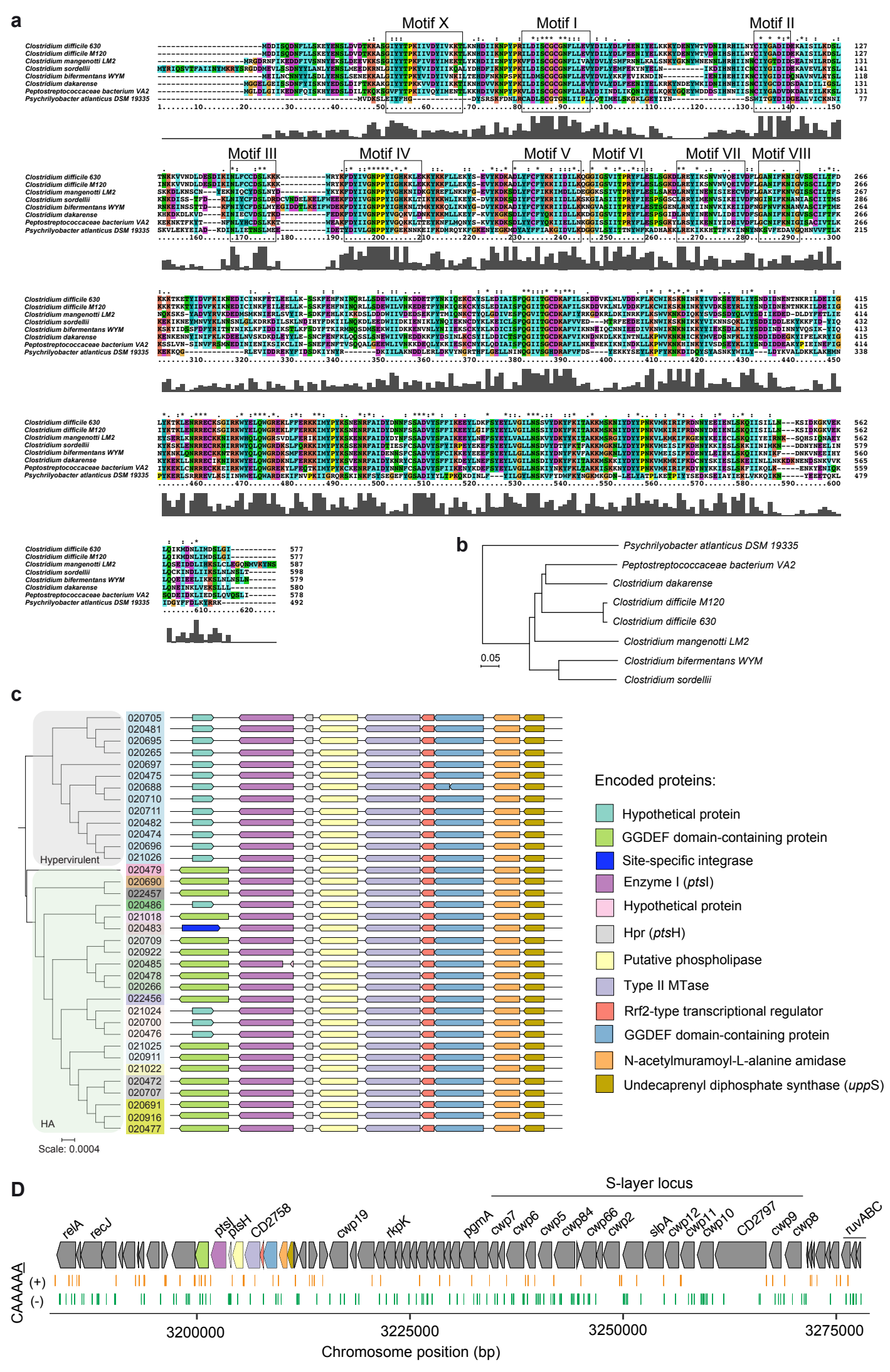

a
Spore amount (a.u.)
150
100
50
0
**
**
WT
ΔCD2758
ΔCD2758-C
ΔCD2758/N165A
b
% H_RES CFUs
6
4
2
0
***
***
WT
ΔCD2758
ΔCD2758-C
ΔCD2758/N165A
c
WT
Δspo0A
ΔCD2758
ΔCD2758-C
ΔCD2758/N165A
H_RES
1.0 ± 0.2
< 10^-7
0.4 ± 0.2
0.7 ± 0.1
0.3 ± 0.1
Phase-dark forespore
Phase-bright forespore
Phase-bright spore
d
CFU / gram feces
10^8
10^6
10^4
10^2
10^0
WT
ΔCD2758-C
ΔCD2758
*
****
*
1
2
3
4
5
6
7
Days post-inoculation

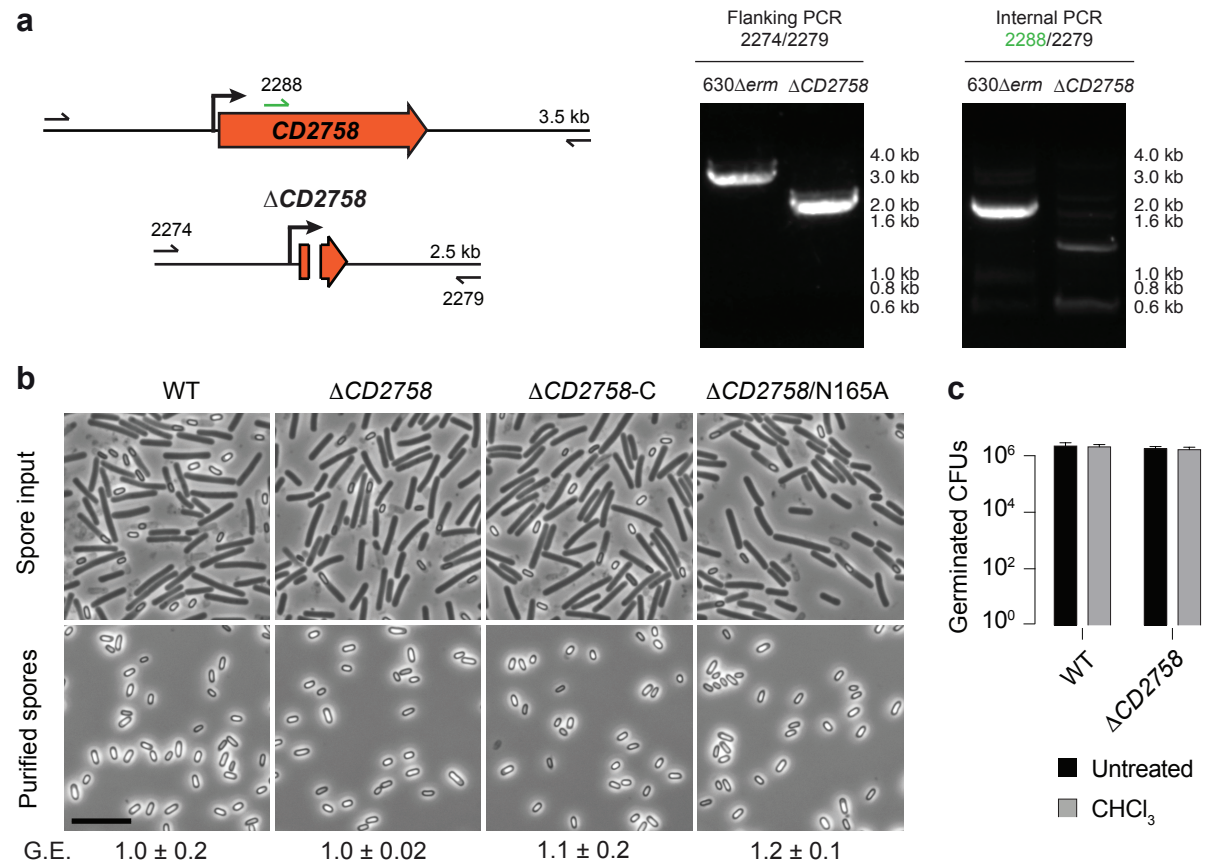

a
MSR Scale
Count
CAAAAA
TTTTTG
Segment length (bp)
Scale
SFC
No SNPs / indels
With SNPs / indels
b
Genome
CAAAAA
Mutated CAAAAA
CDS
Orth. conserved
Orth. variable
Non-orth.
c
Orth. variable
Fold enrichment
Number of genes
CC
Cytoplasm
Flagellum
MF
6P-β-glucosidase
BP
Flagellum, motility

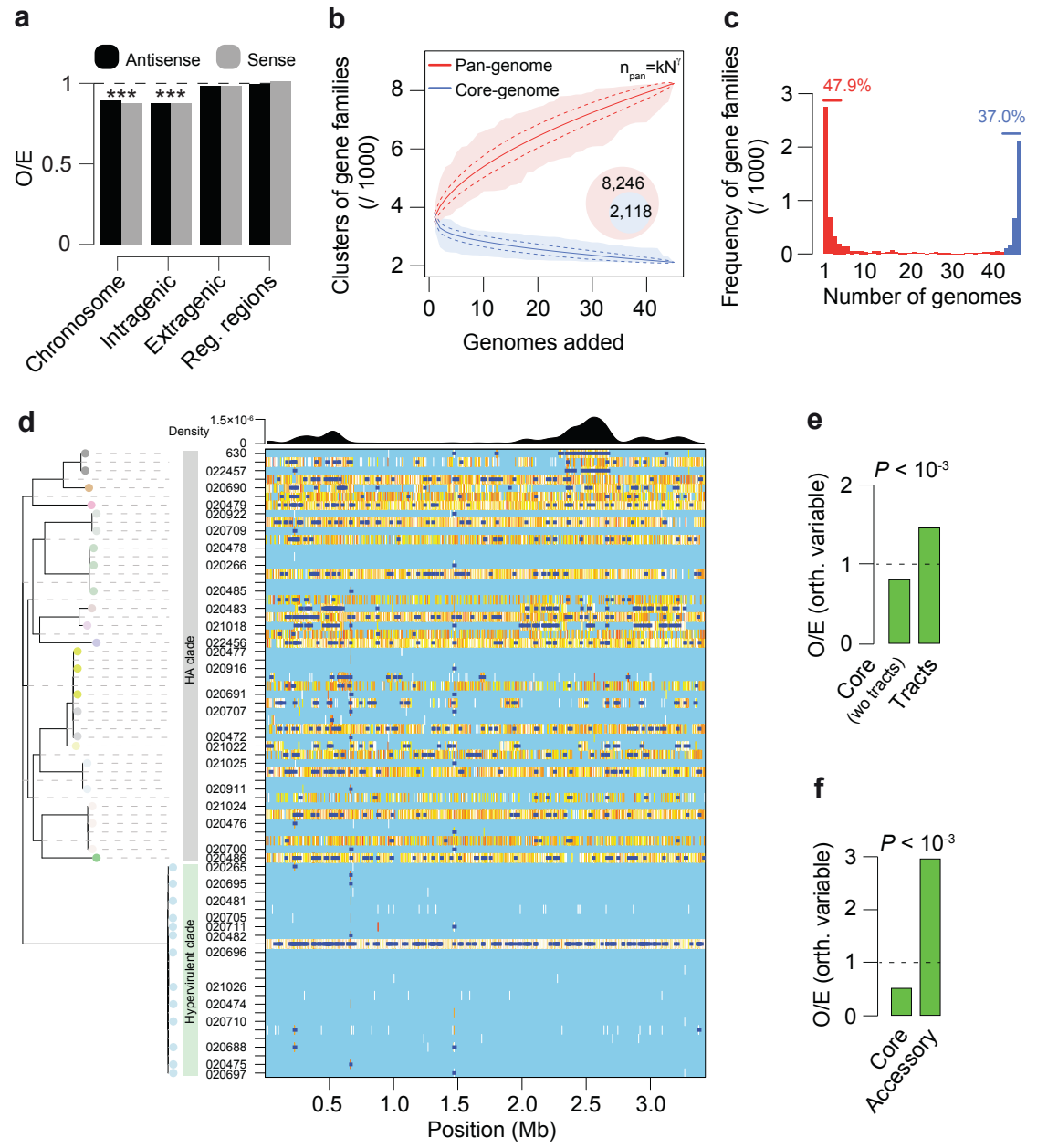

a
Orthologous conserved
Orthologous variable
Non-orthologous
Non-methylated CAAAAA
Number of genomes
CodY
TSS
CodY
CodY
XylR
Chromosome position (Mb)
b
CodY - 786,216
cdu1
tcdB
tcdE
CodY - 2,551,064
CD22010
CD22020
CD22030
CodY - 1,853,965
CD15970
CD15980
XylR - 3,561,672
xylA
xylB
xylR
Chromosome position (kb)
c
CodY
XylR
%CAAAAA overlapping TFBS
NM
M
d
1,316,048
CD11190
CD11200
CD11210
CD11220
Counts / bp
Chromosome position (kb)
e
%CAAAAA overlapping TSS
NM
M
f
CAAAAA positions
Ref
Mutated reads at position i
Total reads at position i
SNP or indel
% mutated reads = Σm_i / Σt_i
CAAAAA
KAAAAA
% mutated reads
MT
EC
CD
HP
Species

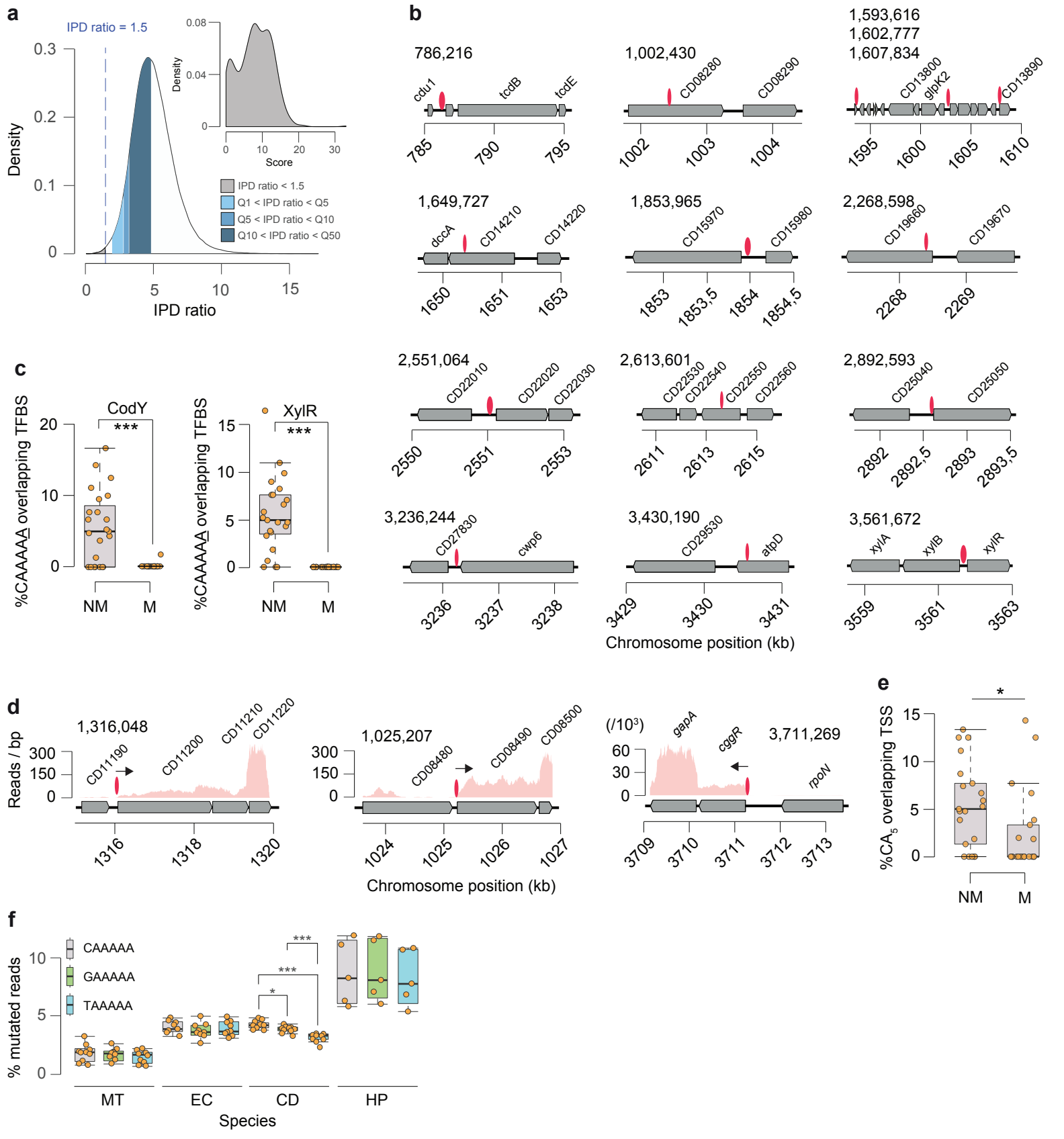

a
Exponential
Sporulation 15 h
Sporulation 19 h
Stationary
Z-score
-2 -1 0 1 2
Fold Enrichment
0 3 6
Bacterial flagellum
Proton trans. (ATP)
Cell division
Proton trans. (ATP)
Bacterial flagellum
Ribosome / Translation
Sporulation
UMP biosynthesis
WT1
WT2
WT3
Mut1
Mut2
Mut3
b
-Log10(P)
0 65 130
This study
1: Exponential
2: Sporulation 15h
3: Sporulation 19h
4: Stationary
Fletcher et al.
5: 24h vs 12h
6: 30h vs 12h
Time post-infection

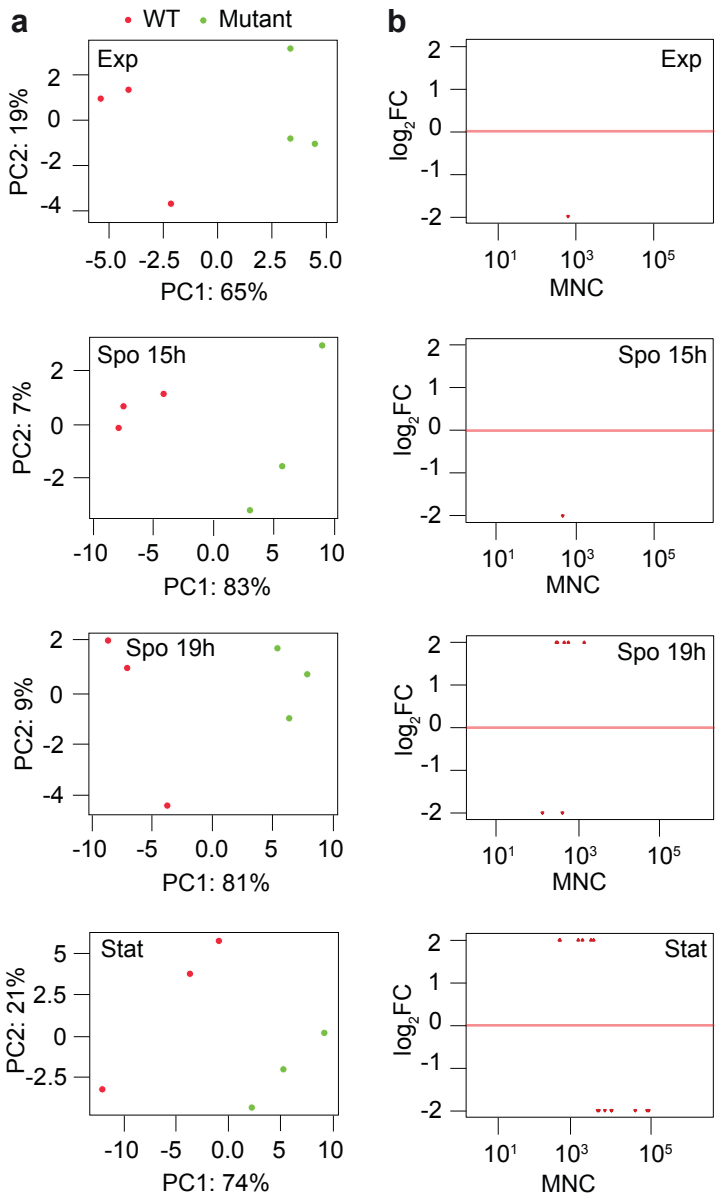

a
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
O / E
4
2
0
***
***
**
**
AGYN5GTCY
ATCN6CTC
ATTYN5CTTC
AYGN6CTG
CAN7TAAAG
GAAN6CTG
GAAN6RTGC
TAAAN4GRTG
TAGN5RTGAA
Type I recognition motif
b
HGT
500
250
0
(g)
(i)
(h)
(e,f)
(a)
ρ = -0.826
P < 0.05
0.2
0.6
1.0
O / E (avg)
c
R-M
Abi
Other
Spacers
T-A
Geneconv
CFML
HGT
020705
020481
020695
020265
020697
020475
020688
020710
020711
020482
020474
020696
021026
020479
020690
022457
020486
021018
020483
020709
020922
020485
020478
020266
022456
021024
020700
020476
021025
020911
021022
020472
020707
020691
020916
020477
φCD111
φCD146
φCD38-2
φCD6356
φMMP04
φCD506
φCDHM11
φCD481-1
φCDHM13
φCD211
φCD27
φMMP02
φCD505
φMMP01
φMMP03
φCDMH1
φC2
φCD119
φCDHM19
S
M - small tail
S
M - long tail
M - medium tail
ST
1
2
4
6
8
10
14
21
42
44
54
55
63
100
110
R-M, Abi, Other
0
1
2
3
Spacers, T-A
0
10
20
HR events
0
40
80
HGT (/100)
0
2.5
5
Phage matches
0
5
10

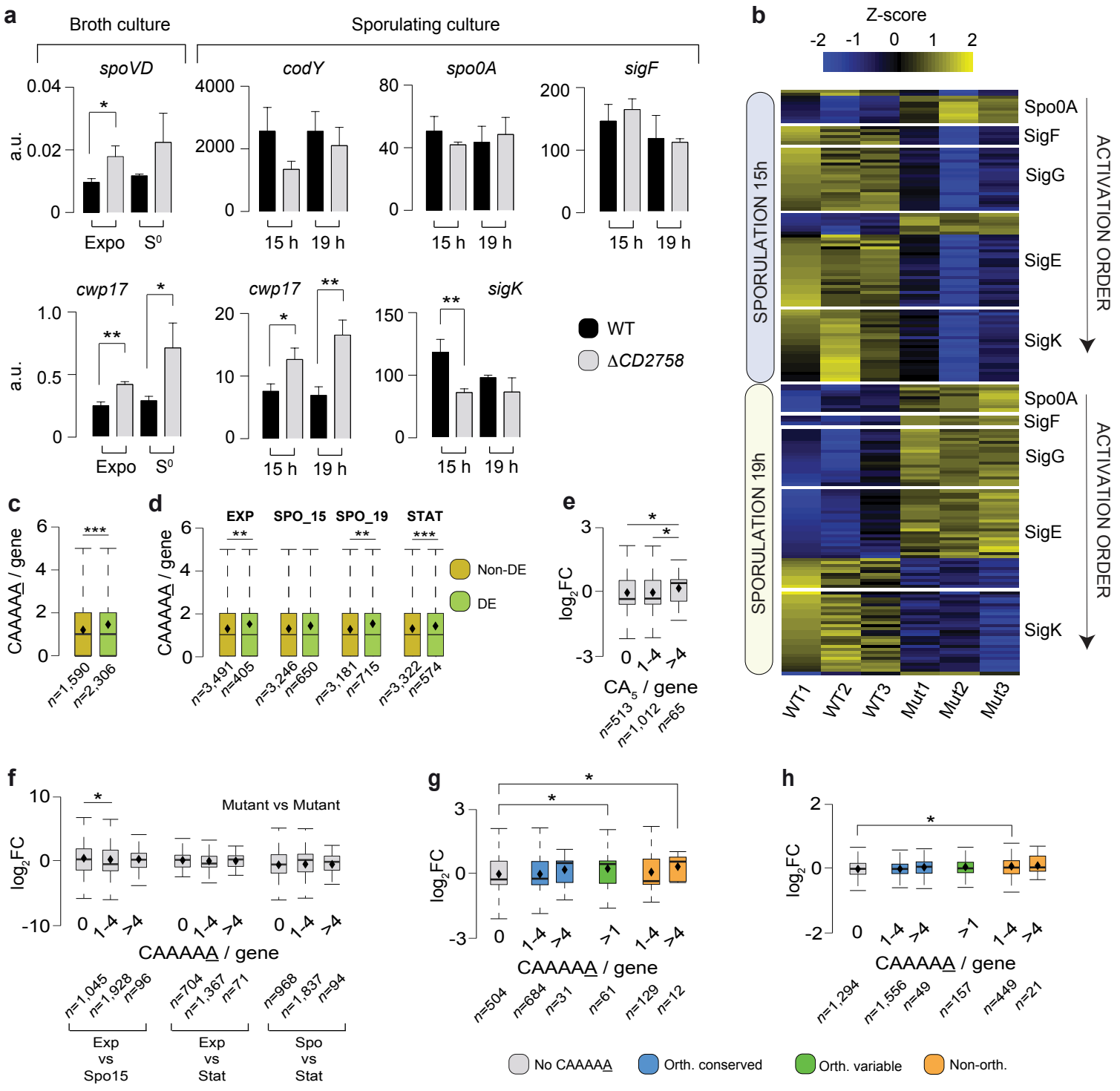

a
Motif X
Motif I
Motif II
Motif III
Motif IV
Motif V
Motif VI
Motif VII
Motif VIII
Clostridium difficile 630
Clostridium difficile M120
Clostridium mangenotti LM2
Clostridium sordellii
Clostridium bifermentans WYM
Clostridium dakarense
Peptostreptococcaceae bacterium VA2
Psychrilyobacter atlanticus DSM 19335
b
Psychrilyobacter atlanticus DSM 19335
Peptostreptococcaceae bacterium VA2
Clostridium dakarense
Clostridium difficile M120
Clostridium difficile 630
Clostridium mangenotti LM2
Clostridium bifermentans WYM
Clostridium sordellii
0.05
c
Hypervirulent
HA
Scale: 0.0004
Encoded proteins:
Hypothetical protein
GGDEF domain-containing protein
Site-specific integrase
Enzyme I (ptsI)
Hypothetical protein
Hpr (ptsH)
Putative phospholipase
Type II MTase
Rrf2-type transcriptional regulator
GGDEF domain-containing protein
N-acetylmuramoyl-L-alanine amidase
Undecaprenyl diphosphate synthase (uppS)
D
S-layer locus
relA
recJ
ptsI
ptsH
CD2758
cwp19
rkpK
pgmA
cwp7
cwp6
cwp5
cwp84
cwp66
cwp2
slpA
cwp12
cwp11
cwp10
CD2797
cwp9
cwp8
ruvABC
CAAAAA
(+)
(-)
3200000
3225000
3250000
3275000
Chromosome position (bp)

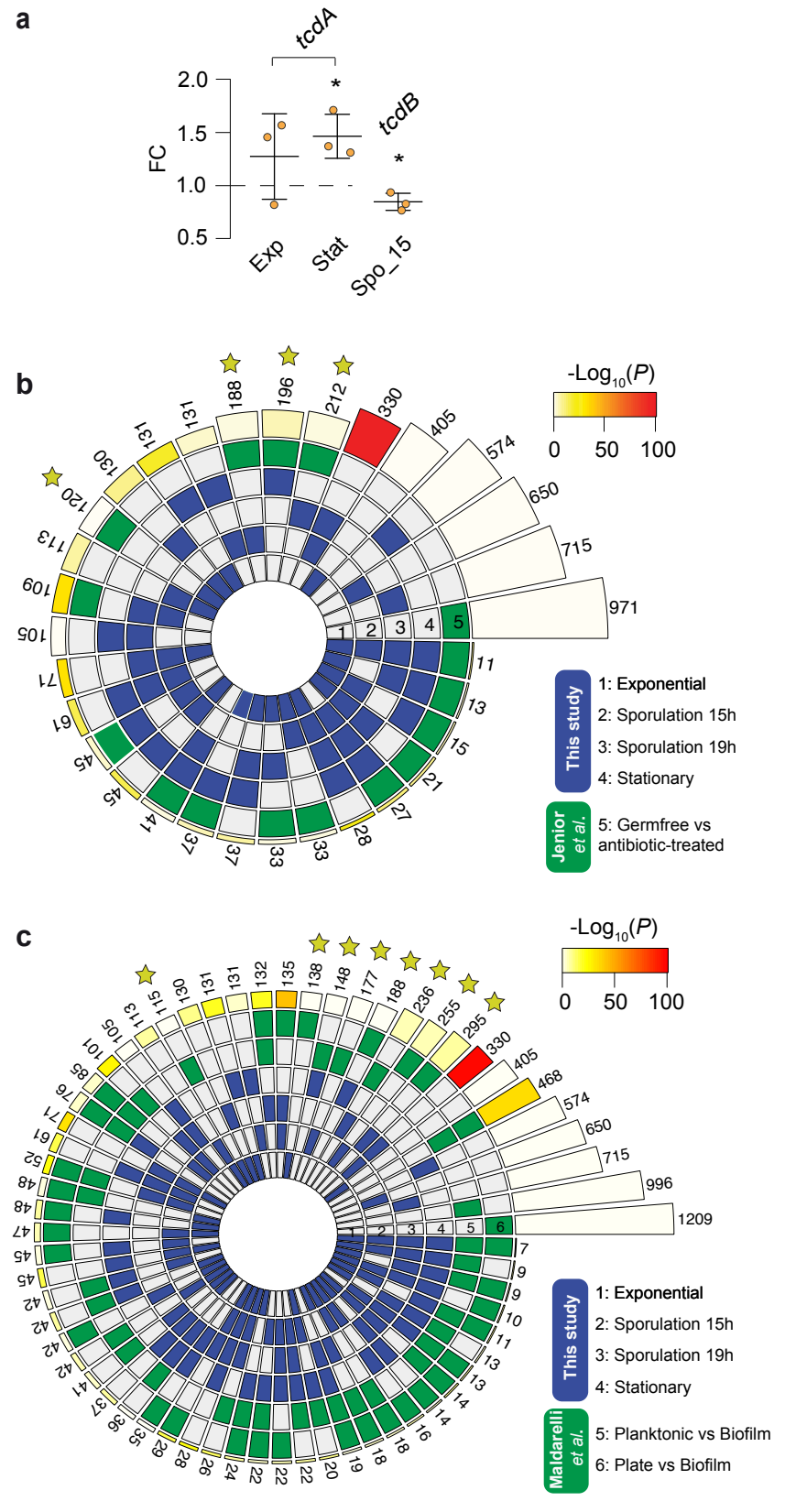

a
2288
CD2758
3.5 kb
ΔCD2758
2274
2.5 kb
2279
Flanking PCR
2274/2279
630Δerm
ΔCD2758
Internal PCR
2288/2279
630Δerm
ΔCD2758
4.0 kb
3.0 kb
2.0 kb
1.6 kb
1.0 kb
0.8 kb
0.6 kb
b
WT
ΔCD2758
ΔCD2758-C
ΔCD2758/N165A
Spore input
Purified spores
G.E.
1.0 ± 0.2
1.0 ± 0.02
1.1 ± 0.2
1.2 ± 0.1
c
Germinated CFUs
10^6
10^4
10^2
10^0
WT
ΔCD2758
Untreated
CHCl3

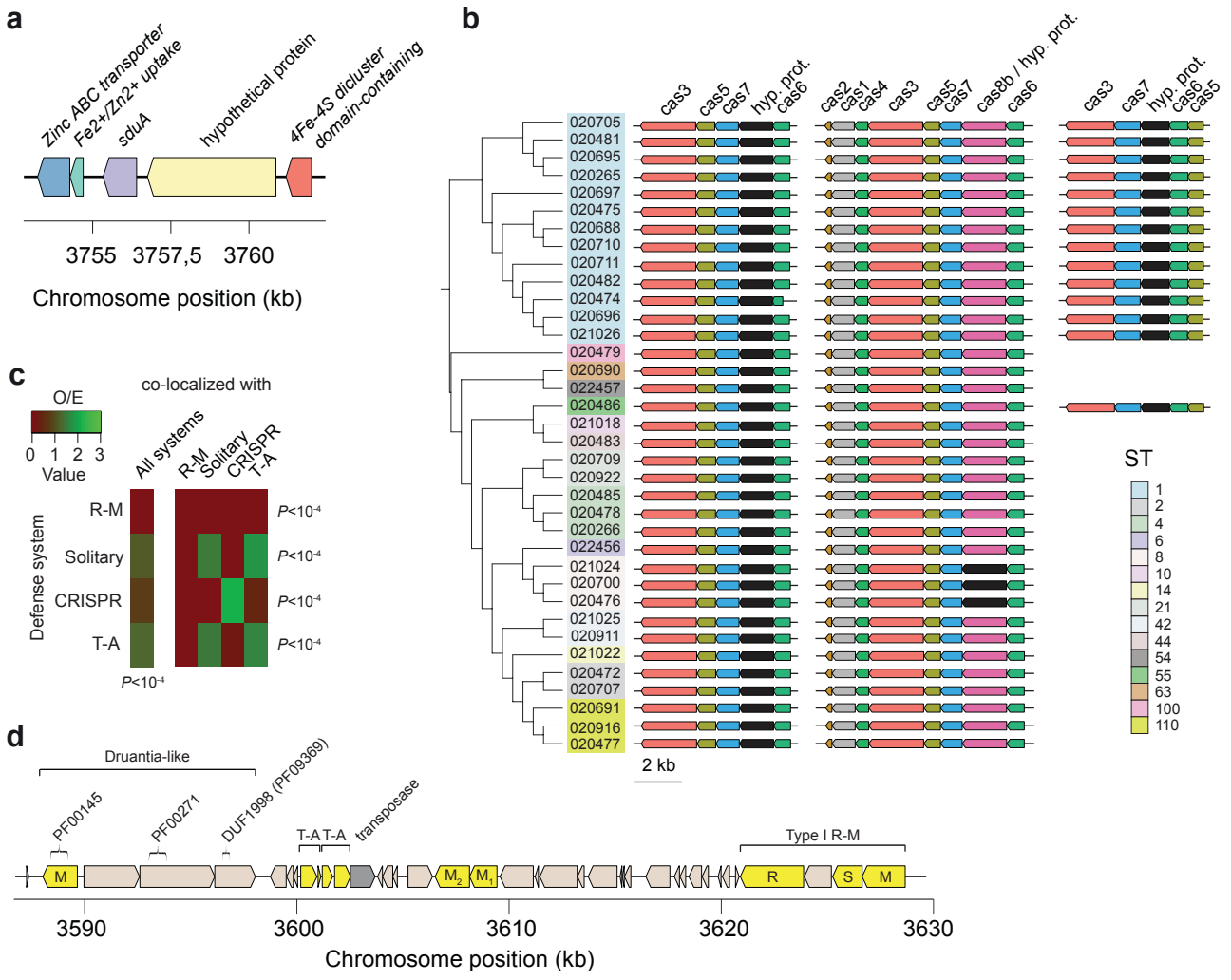

a
Antisense
Sense
O/E
1
0.5
0
***
***
Chromosome
Intragenic
Extragenic
Reg. regions
b
Pan-genome
Core-genome
$n_{pan}=kN^{\gamma}$
8,246
2,118
Clusters of gene families (/ 1000)
Genomes added
c
47.9%
37.0%
Frequency of gene families (/ 1000)
Number of genomes
d
Density
1.5×10-6
HA clade
Hypervirulent clade
Position (Mb)
e
P < 10-3
O/E (orth. variable)
Core (wo tracts)
Tracts
f
P < 10-3
O/E (orth. variable)
Core
Accessory

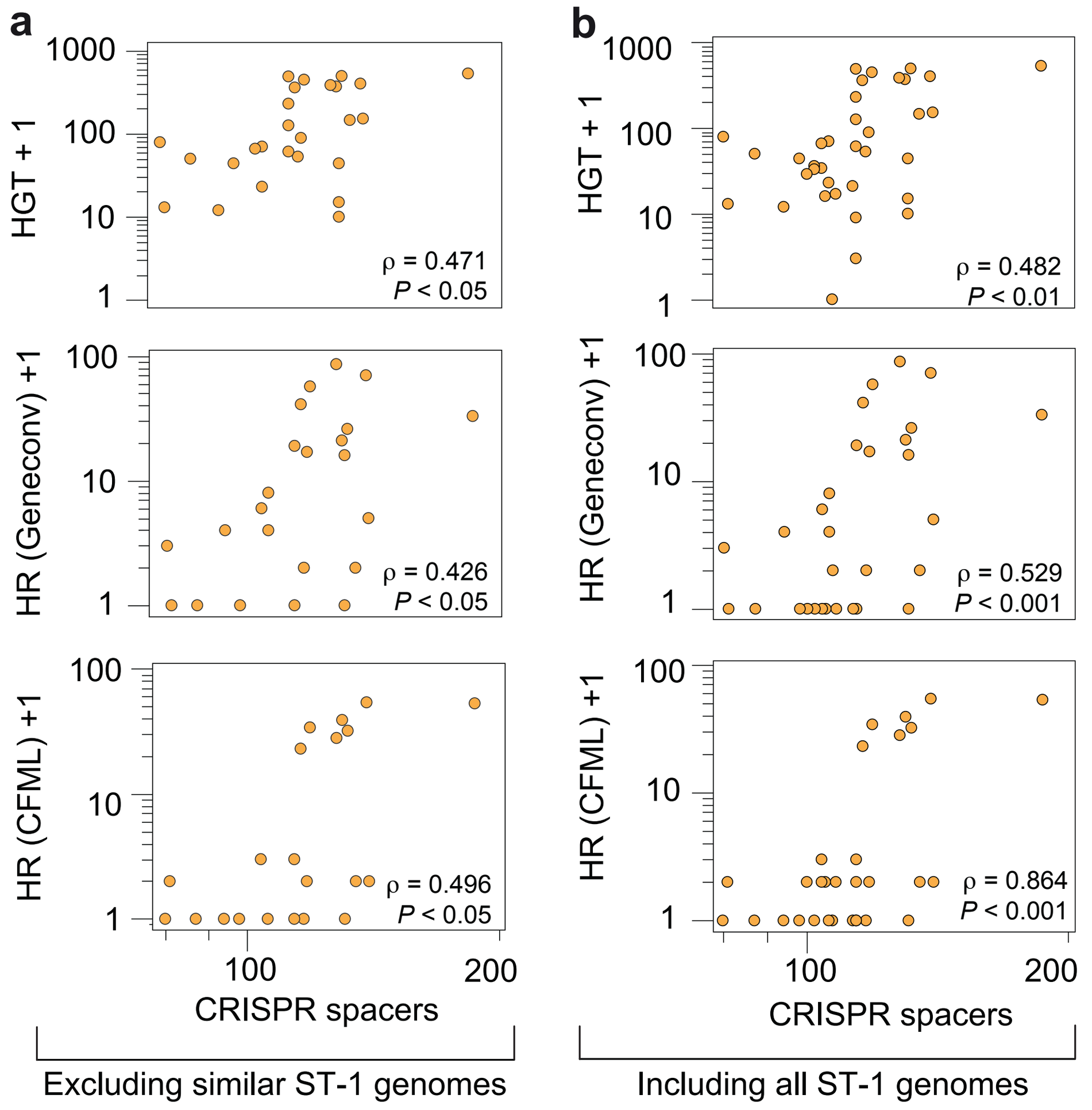

a
IPD ratio = 1.5
Density
IPD ratio
Score
IPD ratio < 1.5
Q1 < IPD ratio < Q5
Q5 < IPD ratio < Q10
Q10 < IPD ratio < Q50
b
786,216
cdu1
tcdB
tcdE
1,002,430
CD08280
CD08290
1,593,616
1,602,777
1,607,834
CD13800
glpK2
CD13890
1,649,727
dccA
CD14210
CD14220
1,853,965
CD15970
CD15980
2,268,598
CD19660
CD19670
2,551,064
CD22010
CD22020
CD22030
2,613,601
CD22530
CD22540
CD22550
CD22560
2,892,593
CD25040
CD25050
3,236,244
CD27830
cwp6
3,430,190
CD29530
atpD
3,561,672
xylA
xylB
xylR
Chromosome position (kb)
c
CodY
XylR
%CAAAAA overlapping TFBS
NM
M
d
1,316,048
CD11190
CD11200
CD11210
CD11220
1,025,207
CD08480
CD08490
CD08500
(/10³)
gapA
cggR
3,711,269
rpoN
Reads / bp
Chromosome position (kb)
e
%CA5 overlapping TSS
NM
M
f
CAAAAA
GAAAAA
TAAAAA
% mutated reads
MT
EC
CD
HP
Species


a
WT
Mutant
Exp
PC2: 19%
PC1: 65%
Spo 15h
PC2: 7%
PC1: 83%
Spo 19h
PC2: 9%
PC1: 81%
Stat
PC2: 21%
PC1: 74%
b
Exp
Spo 15h
Spo 19h
Stat
log2FC
MNC


a
Broth culture
Sporulating culture
spoVD
codY
spo0A
sigF
a.u.
Expo
S0
15 h
19 h
cwp17
sigK
WT
ΔCD2758
b
Z-score
-2 -1 0 1 2
SPORULATION 15h
SPORULATION 19h
Spo0A
SigF
SigG
SigE
SigK
ACTIVATION ORDER
WT1
WT2
WT3
Mut1
Mut2
Mut3
c
CAAAAA / gene
n=1,590
n=2,306
d
EXP
SPO_15
SPO_19
STAT
Non-DE
DE
n=3,491
n=405
n=3,246
n=650
n=3,181
n=715
n=3,322
n=574
e
log2FC
CA5 / gene
n=513
n=1,012
n=65
f
Mutant vs Mutant
n=1,045
n=1,928
n=96
n=704
n=1,367
n=71
n=968
n=1,837
n=94
Exp vs Spo15
Exp vs Stat
Spo vs Stat
g
n=504
n=684
n=31
n=61
n=129
n=12
h
n=1,294
n=1,556
n=49
n=157
n=449
n=21
No CAAAAA
Orth. conserved
Orth. variable
Non-orth.


a
tcdA
tcdB
FC
2.0
1.5
1.0
0.5
Exp
Stat
Spo_15
b
-Log10(P)
0
50
100
This study
1: Exponential
2: Sporulation 15h
3: Sporulation 19h
4: Stationary
Jenior et al.
5: Germfree vs antibiotic-treated
c
-Log10(P)
0
50
100
This study
1: Exponential
2: Sporulation 15h
3: Sporulation 19h
4: Stationary
Maldarelli et al.
5: Planktonic vs Biofilm
6: Plate vs Biofilm


a
Zinc ABC transporter
Fe2+/Zn2+ uptake
sduA
hypothetical protein
4Fe-4S dicluster
domain-containing
3755
3757,5
3760
Chromosome position (kb)
b
cas3
cas5
cas7
hyp. prot.
cas6
cas2
cas1
cas4
cas3
cas5
cas7
cas8b / hyp. prot.
cas6
cas3
cas7
hyp. prot.
cas6
cas5
020705
020481
020695
020265
020697
020475
020688
020710
020711
020482
020474
020696
021026
020479
020690
022457
020486
021018
020483
020709
020922
020485
020478
020266
022456
021024
020700
020476
021025
020911
021022
020472
020707
020691
020916
020477
2 kb
ST
1
2
4
6
8
10
14
21
42
44
54
55
63
100
110
c
O/E
0 1 2 3
Value
co-localized with
All systems
R-M
Solitary
CRISPR
T-A
Defense system
R-M
Solitary
CRISPR
T-A
P<10-4
P<10-4
P<10-4
P<10-4
P<10-4
d
Druantia-like
PF00145
PF00271
DUF1998 (PF09369)
T-A
T-A
transposase
Type I R-M
M
M2
M1
R
S
M
3590
3600
3610
3620
3630
Chromosome position (kb)


a
b
HGT + 1
HR (Geneconv) +1
HR (CFML) +1
1000
100
10
1
ρ = 0.471
P < 0.05
ρ = 0.426
P < 0.05
ρ = 0.496
P < 0.05
ρ = 0.482
P < 0.01
ρ = 0.529
P < 0.001
ρ = 0.864
P < 0.001
100
200
CRISPR spacers
Excluding similar ST-1 genomes
Including all ST-1 genomes