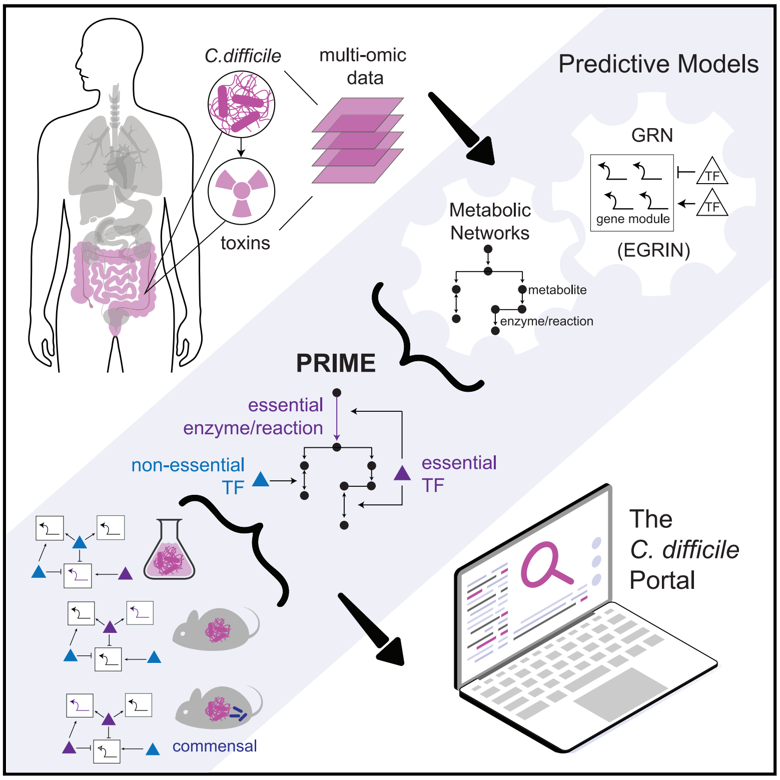
Resource

Cell Host & Microbe

Predictive regulatory and metabolic network models for systems analysis of *Clostridioides difficile*

Graphical abstract



Highlights

- The EGRIN model uncovers regulatory responses of *C. difficile* in multiple contexts
- Genes and pathways needed to support *C. difficile* growth *in vivo* are identified
- PRIME predicts *in vivo* synergistic epistasis between transcription factor networks
- The *C. difficile* portal makes all tools and resources available to the public

Authors

Mario L. Arrieta-Ortiz,
Selva Rupa Christinal Immanuel,
Serdar Turkarslan, ..., Bruno Dupuy,
Lynn Bry, Nitin S. Baliga

Correspondence

nitin.baliga@isbscience.org

In brief

C. difficile is one of the leading causes of hospital-acquired infections. Arrieta-Ortiz et al. report three predictive models (with extensive validations) for dissecting interplay of regulation and metabolism that underlies host-pathogen interactions of *C. difficile*. The *C. difficile* interactive web portal provides access to these models, compiled datasets, and algorithms.

Arrieta-Ortiz et al., 2021, Cell Host & Microbe *29*, 1–15
November 10, 2021 Published by Elsevier Inc.
https://doi.org/10.1016/j.chom.2021.09.008

CellPress

Please cite this article in press as: Arrieta-Ortiz et al., Predictive regulatory and metabolic network models for systems analysis of *Clostridioides difficile*, Cell Host & Microbe (2021), https://doi.org/10.1016/j.chom.2021.09.008

Cell Host & Microbe

CellPress

Resource

Predictive regulatory and metabolic network models for systems analysis of *Clostridioides difficile*

Mario L. Arrieta-Ortiz,[1] Selva Rupa Christinal Immanuel,[1] Serdar Turkarslan,[1] Wei-Ju Wu,[1] Brintha P. Girinathan,[2,5] Jay N. Worley,[2] Nicholas DiBenedetto,[2,6] Olga Soutourina,[3] Johann Peltier,[3] Bruno Dupuy,[4] Lynn Bry,[2] and Nitin S. Baliga[1,7,*]

[1] Institute for Systems Biology, Seattle, WA 98109, USA
[2] Massachusetts Host-Microbiome Center, Brigham & Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
[3] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-yvette 91198, France
[4] Laboratoire Pathogenèse des Bactéries anaérobies, Institut Pasteur, Université de Paris, UMR CNRS 2001, Paris 75015, France
[5] Present address: Ginkgo Bioworks, Boston, MA 02210, USA
[6] Present address: Department of Molecular Biology and Microbiology, Tufts University School of Medicine, Boston, MA 02111, USA
[7] Lead contact
*Correspondence: nitin.baliga@isbscience.org

https://doi.org/10.1016/j.chom.2021.09.008

SUMMARY

We present predictive models for comprehensive systems analysis of *Clostridioides difficile*, the etiology of pseudomembranous colitis. By leveraging 151 published transcriptomes, we generated an EGRIN model that organizes 90% of *C. difficile* genes into a transcriptional regulatory network of 297 co-regulated modules, implicating genes in sporulation, carbohydrate transport, and metabolism. By advancing a metabolic model through addition and curation of metabolic reactions including nutrient uptake, we discovered 14 amino acids, diverse carbohydrates, and 10 metabolic genes as essential for *C. difficile* growth in the intestinal environment. Finally, we developed a PRIME model to uncover how EGRIN-inferred combinatorial gene regulation by transcription factors, such as CcpA and CodY, modulates essential metabolic processes to enable *C. difficile* growth relative to commensal colonization. The *C. difficile* interactive web portal provides access to these model resources to support collaborative systems-level studies of context-specific virulence mechanisms in *C. difficile*.

INTRODUCTION

Clostridioides difficile, the etiology of pseudomembranous colitis, causes more than 500,000 infections, 30,000 deaths, and $5 billion per year in US healthcare costs (Lessa et al., 2015). Infections arise through a variety of conditions that modulate the pathogen's ability to colonize and expand in the gut. Antibiotic ablation of the commensal microbiota alters nutrient states in intestinal environments due to lack of competition for nutrients from host, dietary, or microbial origin. The pathogen modifies its metabolism to respond to these altered states, which stimulates subsequent cellular programs that can promote colonization and growth. Stress and starvation conditions trigger *C. difficile* sporulation, biofilm formation, and release of mucosal damaging toxins (Aktories, 2011; Antunes et al., 2012; Saujet et al., 2011).

Symptomatic infection requires the production of toxins from the *C. difficile* pathogenicity locus (PaLoc), which includes the genes *tcdA, tcdB*, and *tcdE* that respectively encode the A and B toxins and holin involved in toxin export (Govind and Dupuy, 2012). The PaLoc also contains *tcdR* and *tcdC* sigma and anti-sigma factors, respectively (Mani and Dupuy, 2001; Matamouros et al., 2007). *C. difficile* elaborates toxin to extract nutrients from the host and promote spore shedding (Edwards et al., 2016a; Martin-Verstraete et al., 2016; Walter et al., 2014). Regulation of PaLoc expression occurs via a complex network of transcription factors (TFs) and small molecule inputs, of which direct primary regulators have been described, but more complex and combinatorial effects remain unclear (Martin-Verstraete et al., 2016). Toxin production triggers host immune responses that alter the redox state of the gut environment and can induce *C. difficile* stress responses to cell wall, oxidative, and other damaging stimuli (Bradshaw et al., 2017; Kint et al., 2017; Neumann-Schaal et al., 2018; Woods et al., 2016). As per all microbes, *C. difficile* adapts to complex, dynamic environments through changes in metabolism coordinated by a gene regulatory network (GRN) (Brooks et al., 2011; Elena and Lenski, 2003). However, the mechanisms by which the GRN and metabolic pathways integrate to modulate *C. difficile* pathogenesis remain ill-defined (McDonald et al., 2018; Vemuri et al., 2017).

The *C. difficile* 630 (CD630) genome encodes 4,018 genes, with ~309 candidate TFs (including sigma factors), 1,030 metabolic genes, and 1,330 genes with unknown function (Monot et al., 2011; Riedel et al., 2015). The clinical ATCC43255 strain of *C. difficile*, used to capitulate symptomatic infections in mouse models, encodes 4,117 genes and ~327 putative TFs, of which ~97% have orthologs in the CD630 strain (Girinathan et al., 2021). To address questions regarding the broader systems-level interplay among genes in colonization and infection, we used computational modeling and network inference algorithms to construct an environment and gene regulatory influence network (EGRIN) model for *C. difficile*. This model leverages a compendium of 151 public transcriptomes that surveyed responses of CD630 in diverse contexts. The EGRIN model consists of modules of putatively co-regulated genes identified based on their co-expression over subsets of conditions, enrichment of functional associations, chromosomal proximity, and shared *cis*-acting gene regulatory elements (GREs) within their promoter regions. Further, using regression analysis, EGRIN also captures the combinatorial regulation of genes within each module as a function of the weighted influences of TFs. The model supports a systems-level understanding of the infective capacity of this obligate anaerobe under different *in vitro* and *in vivo* conditions.

In addition, we have advanced a metabolic network model of *C. difficile* to understand how conditional regulation manifests physiologically, by adding reactions and associated genes supporting the exchange of nutrients required for growth within the host. Integration of transcriptional and metabolic networks into a phenotype of regulatory influences integrated with metabolism and environment (PRIME) model supports prediction of conditional fitness contribution of every TF and metabolic gene of *C. difficile* (Immanuel et al., 2021). Analyses uncover TFs driving essential adaptive responses *in vivo*. This analytic framework provides a systems-level view of the transcriptional and metabolic networks that coordinate *C. difficile*'s colonization, growth, expression of toxin, and adaptions to changing environments with host infection. Our models identified multiple TFs that coordinate critical aspects within each of these components, including contributions from PrdR, which regulates the Stickland proline and glycine reductase systems and other energy-generating pathways, and Rex, a regulator modulating energy balance in *C. difficile* (Bouillaut et al., 2013, 2019). We successfully validated PRIME-identified enhanced epistasis between *ccpA* and *codY* in the presence of a protective gut commensal. These findings refine the context and roles of these and other regulators in *C. difficile* virulence and provide specific targets of vulnerability for model-informed interventions against this pathogen. The compiled datasets, algorithms, and models can be explored interactively through a community-wide web resource at http://networks.systemsbiology.net/cdiff-portal/.

RESULTS

Reconstruction of the environment and gene regulatory influence network (EGRIN) model for *C. difficile* 630

To investigate *C. difficile*'s transcriptionally driven adaptive strategies, we compiled 151 public transcriptomes from 11 independent studies on CD630 (Table S1). This compendium captures diverse transcriptional responses of *C. difficile* to commensals, *in vitro* and *in vivo* responses to nutrient conditions, and consequences of TF deletions. The transcriptome compendium together with functional associations information was analyzed with a suite of network inference tools (i.e., cMonkey2 and the Inferelator) to infer an EGRIN model for *C. difficile* (Figure 1A; Arrieta-Ortiz et al., 2015; Reiss et al., 2015). The resulting EGRIN model organized 3,995 of 4,018 CD630 genes into 406 gene modules and inferred module regulation by 138 of 309 genomically identified TFs that putatively act through GREs discovered within gene and operon promoters. Among the Inferelator implicated regulatory networks, 255 modules were controlled by more than one TF, and 120 were regulated by more than two TFs (Figure S1). The TF module assignments support subsequent hypothesis-driven design of ChIP-seq and TF-deletion experiments to validate the regulatory network architecture under physiologically relevant environments.

Residual scores reflect the coherence of gene co-expression patterns and evaluated the quality of modules within the EGRIN model (Reiss et al., 2006). The lower the residual score, the higher the quality of the module. Using an empirical approach, a residual cutoff of 0.55 identified a functionally meaningful set of 297 high-quality modules (73% of the total 406 modules) based on the relative enrichment of related functions within modules that passed filtering (Figure 1B). This threshold was similar to the threshold used to identify high-quality EGRIN modules for *Mycobacterium tuberculosis* (Peterson et al., 2014). The high-quality modules captured transcriptional regulation of 3,617 genes (90%) in CD630, with average membership of 20 genes per module (Figures 1C and 1D). These metrics were consistent with models developed for other organisms (Brooks et al., 2014; Peterson et al., 2014), a remarkable finding given that the transcriptional dataset used to construct the *C. difficile* model was less than 10% the size of compendia used to construct models for other species.

Validation of the modular architecture and regulatory mechanisms uncovered by the *C. difficile* EGRIN model

We tested the accuracy of the EGRIN model to reconstruct previously characterized regulons and recapitulate key aspects of *C. difficile* biology. We performed functional enrichment analysis using an updated annotation of *C. difficile* genome (Girinathan et al., 2021). This analysis identified 93 of 297 modules (31%) with significant enrichment of functionally related genes in 45 pathways (hypergeometric test adjusted p value $\leq$ 0.05). Among these pathways, 14 were overrepresented in three or more modules (Figure 2A), highlighting the capacity of the model to discover conditional partitioning of cellular processes. We also investigated whether the EGRIN model identified known regulatory interactions between TFs and their target genes. We compiled regulons (i.e., target genes) of 13 characterized TFs in *C. difficile*, representing a network of 1,349 TF-gene interactions (Table S2). Notably, a total of 57 modules (19% of all high-quality modules) were significantly enriched with nine of these TF regulons (Figure 2B). The EGRIN model recapitulated 514 of the 1,208 (42.5%) previously characterized interactions, a value consistent with the EGRIN recall rate for *M. tuberculosis* (41%–49%) (Peterson et al., 2014). The poor recall of the remaining four regulons (141 regulatory interactions) could be due to underrepresentation of expression data from conditions in which these regulons are active. This analysis also uncovered combinatorial regulation of genes across 19 modules (i.e., enriched with more than one TF regulon). Consistent with the known hierarchical scheme for regulation of sporulation (Saujet et al., 2013), Spo0A putatively influenced expression of 161 genes across at least eight modules in combination with

Please cite this article in press as: Arrieta-Ortiz et al., Predictive regulatory and metabolic network models for systems analysis of *Clostridioides difficile*, Cell Host & Microbe (2021), https://doi.org/10.1016/j.chom.2021.09.008

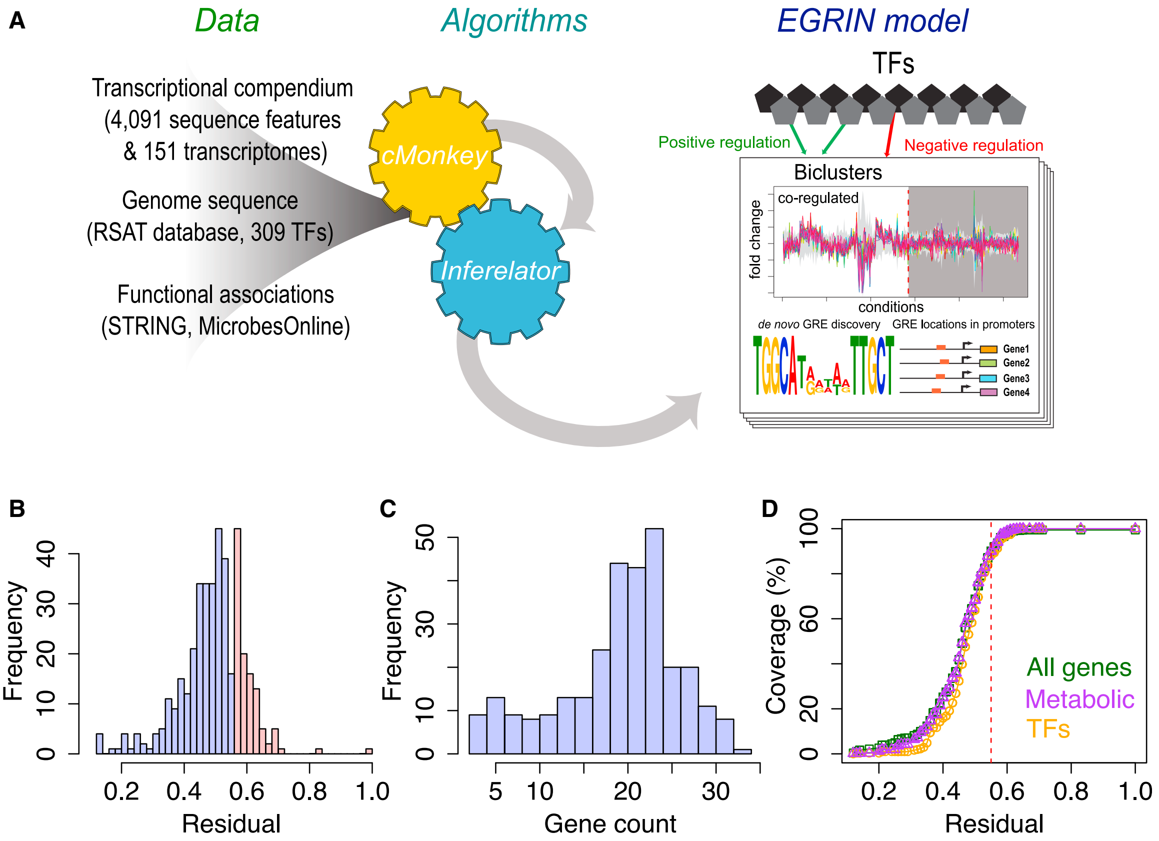


Figure 1. Inference pipeline and general properties of the resulting EGRIN model of *C. difficile*

(A) Framework used to build the EGRIN model.

(B) Distribution of residual values for the 406 detected co-regulated gene modules. 297 modules with residual ≤ 0.55 (shown in purple) were labeled as high quality.

(C) Distribution of gene count for the high-quality gene modules.

(D) Coverage of all genes (4,018), the subset of metabolic genes (1,030), and TFs (309) by EGRIN modules for different residual thresholds. The red dashed line indicates the 0.55 residual cutoff.

sigma factors implicated in sporulation (e.g., SigE). EGRIN also predicted CcpA contributions in six modules in combination with CodY, PrdR, and SigL, illustrating the complexity of modular transcriptional regulation in *C. difficile*. EGRIN detected gene co-regulation within and across functionally related operons. For example, module #152, enriched with the SigD regulon, contains 16 genes among four operons including the flagellar operon *flgG1G-fliMN*-CD630_02720-*htpG*, and *pyrBKDE*, CD630_30270-CD630_30280-*malY*-CD630_30300, and CD630_32430-*prdA* (Figure S2A).

Gene clustering in EGRIN is constrained by the *de novo* discovery of conserved GRE(s) within promoters to cluster genes that are co-regulated, and not just co-expressed. The GREs represent putative TF-binding sites that are independently implicated by the Inferelator and protein-DNA interaction maps as regulators of genes within the same module (Bonneau et al., 2007; Brooks et al., 2014; Peterson et al., 2014). For instance, Inferelator regression-based analysis assigned 138 TFs as putative regulators of the EGRIN modules, hypothesizing TF-GRE associations. While the paucity of characterized binding sites in *C. difficile* limited validation of TF-GRE mappings, the predicted context-specific TF regulation of genes within modules can guide ChIP-seq experiments to support validation. Notably, we determined that the GREs within promoters of genes in modules #182 and #308 recapitulated the previously characterized consensus binding sequences for CodY and SigL (Figures 2C, 2D, S2B, and S2C) (Dineen et al., 2007; Soutourina et al., 2020).

C. difficile's EGRIN model uncovers regulatory networks for the pathogenicity locus

We evaluated EGRIN capacity to recall known mechanisms of PaLoc regulation and provide additional information regarding complex regulatory effects on toxin production. The EGRIN model captured certain previously described effects of CodY on toxin gene expression (Figure 2E), as shown in module #182, which is enriched with CodY targets including *tcdA*. In agreement with the EGRIN-predicted CodY regulation of PaLoc genes, genes encoding the toxin *tcdA* and its regulator *tcdR* were significantly overexpressed upon deletion of *codY* (Figure 2G). Interestingly, *tcdB* (which is not part of module #182) was also upregulated in the *codY* mutant, suggesting that this effect might be an indirect consequence of disrupted CodY regulation of *tcdR* (Figure 2G). Lower affinity of CodY for the *tcdA* promoter has been proposed (Dineen et al., 2007).

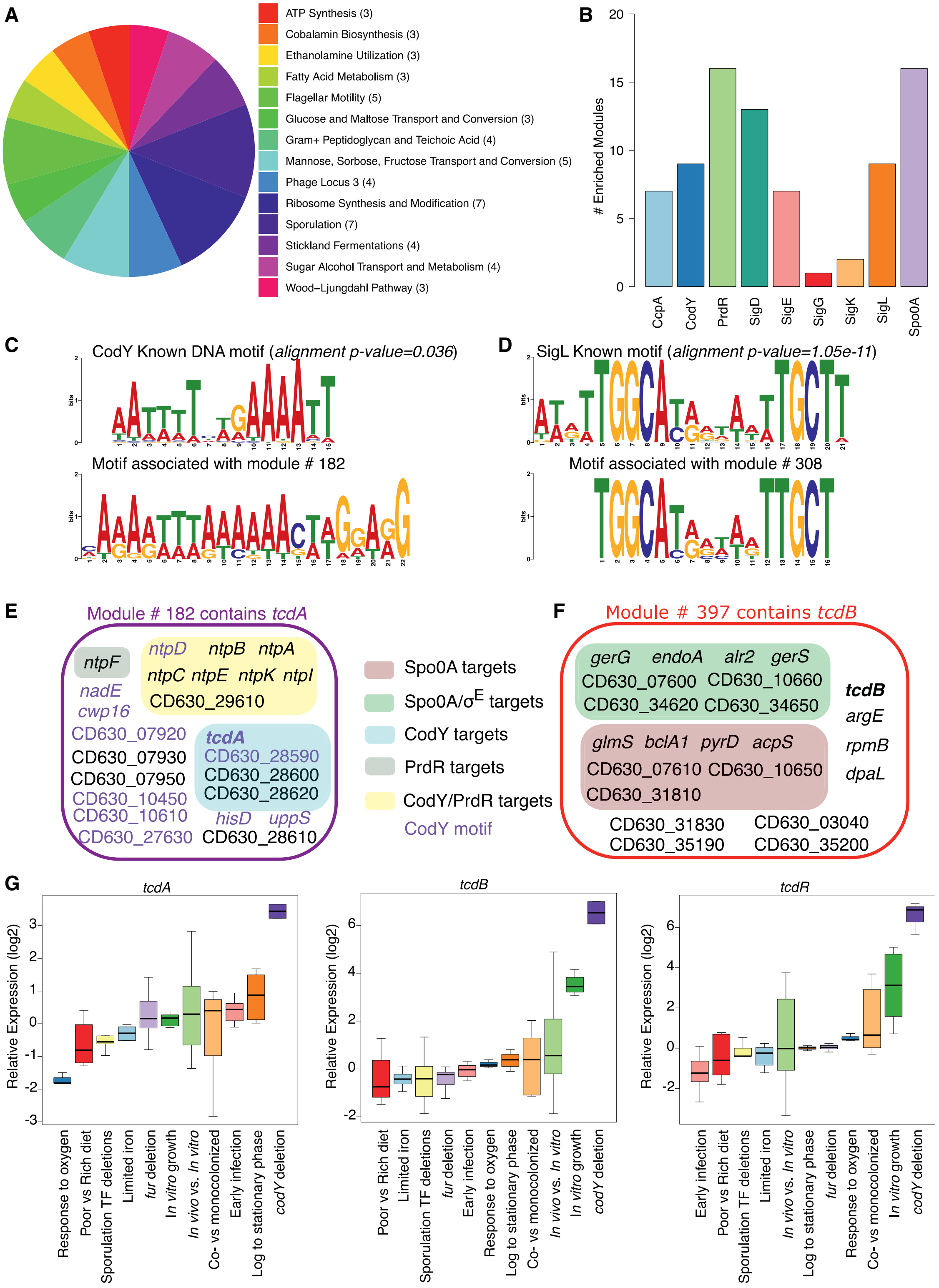


Figure 2. The EGRIN model of *C. difficile* recapitulates known biology of the pathogen

(A) Co-regulated gene modules are enriched with functional terms derived from expert curated annotation of the *C. difficile* genome (Girinathan et al., 2021). The pie chart shows terms over-represented in three or more modules. Number of modules associated with each functional term is shown in parenthesis.

(B) Enriched EGRIN modules among nine (out of 13) manually defined and experimentally supported TF regulons (Table S2).

(C) EGRIN identified the known DNA binding motif of CodY (Dineen et al., 2010).

(legend continued on next page)

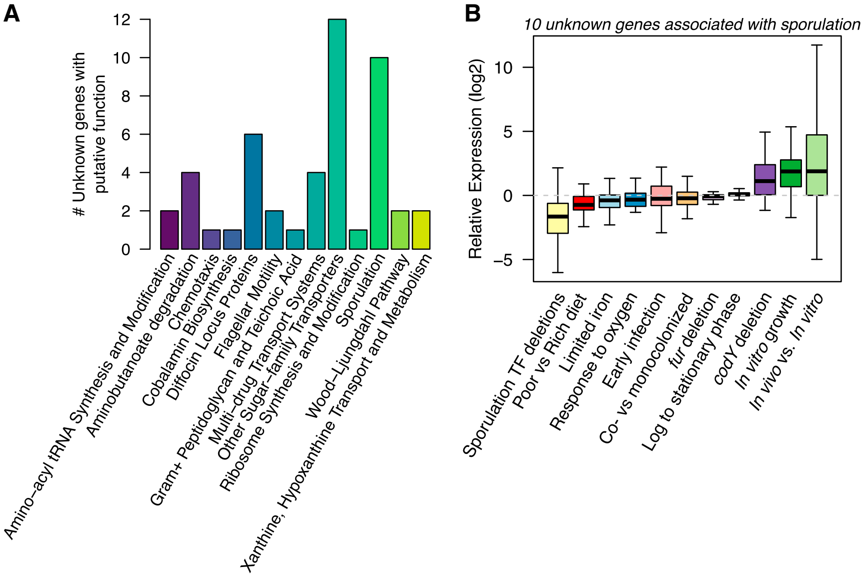


Figure 3. The EGRIN model offers insights on potential functions of uncharacterized genes of *C. difficile*

Hypotheses regarding the functions of 48 uncharacterized genes were generated based on their membership in high-quality EGRIN modules significantly enriched with specific functional terms.

(A) Barplot with the number of unknown genes associated with each functional term (from the *C. difficile* genome annotation in Girinathan et al., 2021).

(B) The involvement of 10 uncharacterized genes in sporulation was supported by their significant downregulation (with respect to a wild-type control) in single deletion strains of sporulation regulators (Table S1). Boxes cover the 25th–75th percentile range (median is indicated by horizontal black line).

However, the presence of the CodY motif in most members of module #182, including *tcdA* (purple font in Figure 2E) suggests direct influence of CodY on *tcdA*. The EGRIN model also identified previously reported connections between sporulation and toxin production (Underwood et al., 2009). The *tcdB* toxin gene was assigned to module #397, which was significantly enriched with genes controlled by Spo0A, the master regulator of sporulation (Figure 2F). Additional members of the PaLoc were assigned to other modules, supporting the presence of multiple condition-dependent promoters within the PaLoc (Table S3).

Assignment of putative functions to genes in EGRIN modules

Approximately 33% of gene features in the CD630 genome have unknown functions. The *C. difficile* EGRIN model provides a resource to define putative functions of uncharacterized genes per functional associations among co-regulated genes (i.e., guilt-by-association) (Wolfe et al., 2005). We predicted functions for 48 uncharacterized genes by mining functional enrichment of modules under different experimental conditions (Table S4), involving 13 functional categories such as "sporulation" and "other sugar-family transporters" (Figure 3A).

Ten genes were assigned sporulation-related functions based on their presence in the modules #206 and #251 (Figure 3B). Module #206 includes 7 stage-III sporulation genes and 2 stage-IV sporulation genes (Figure S3A). Module #251 includes the sporulation-associated sigma factors SigG and SigE (located in the same operon) (Figure S3B). Decreased expression of the 10 genes upon deletion of sporulation sigma factors supported their role in sporulation. Seven genes associated with mother cell-specific roles based on their decreased expression in *sigE* (six genes) and *sigK* (one gene) deletion strains. Two additional genes were downregulated in a *sigG* deletion strain, suggesting functions in the forespore. Notably, Tn-seq studies for gene essentiality in *C. difficile* identified 7 of the 10 genes as required for sporulation (Dembek et al., 2015) (Figures S3A and S3B).

Module #48 contains two adjacent operons (*4hbd-cat2*-CD630_23400-*abfD* and *sucD-cat1*) associated with aminobutanoate degradation (Figure S3C). CodY and PrdR regulate both operons. Hence, we predicted that the 4 uncharacterized genes in this module may also support amino acid metabolism (Table S4). In support of this hypothesis, CD630_08760 and CD630_08780 are both differentially expressed upon *codY* deletion. Recent studies suggest that CD630_ 08760 may function as a tyrosine transporter per its homology to the CodY-regulated neighbor gene, CD630_08730 (Bradshaw et al., 2019). Decreases in tyrosine uptake and Stickland fermentation in clinical isolates lacking CD630_08760 and CD630_08780 further support this hypothesis (Steglich et al., 2018). Altogether, we assigned 48 genes to functions in sporulation (10 genes), sugar transport (12 genes), and others. These examples illustrate use of context-specific co-regulation of genes in EGRIN to predict functions for uncharacterized genes. EGRIN can also aid in designing experiments (e.g., gene perturbations) to validate these predictions within relevant environmental contexts.

EGRIN uncovers differentially active regulatory networks during *in vivo* infection

We investigated the differential activity of EGRIN modules across conditions (Table S1) to discover regulatory mechanisms

(D) EGRIN also identified the known DNA binding motif of SigL (Soutourina et al., 2020).

(E) The EGRIN model recapitulated the previously reported influence of CodY on *tcdA* expression. The module #182 contains *tcdA*, it is enriched with the CodY regulon and contains a GRE (shown in C) similar to the experimentally determined CodY motif.

(F) The EGRIN model captured the interaction between toxin expression and sporulation via module #397 that contains *tcdB* and is enriched with genes regulated by sporulation-related transcriptional regulators.

(G) Expression profiles of *tcdAB* and *tcdR* (positive regulator of the PaLoc). Log_2 ratios were computed with respect to dataset-specific references (Table S1) and grouped by condition blocks. Boxes cover the 25th–75th percentile range (median indicated by horizontal black line).

Please cite this article in press as: Arrieta-Ortiz et al., Predictive regulatory and metabolic network models for systems analysis of *Clostridioides difficile*, Cell Host & Microbe (2021), https://doi.org/10.1016/j.chom.2021.09.008

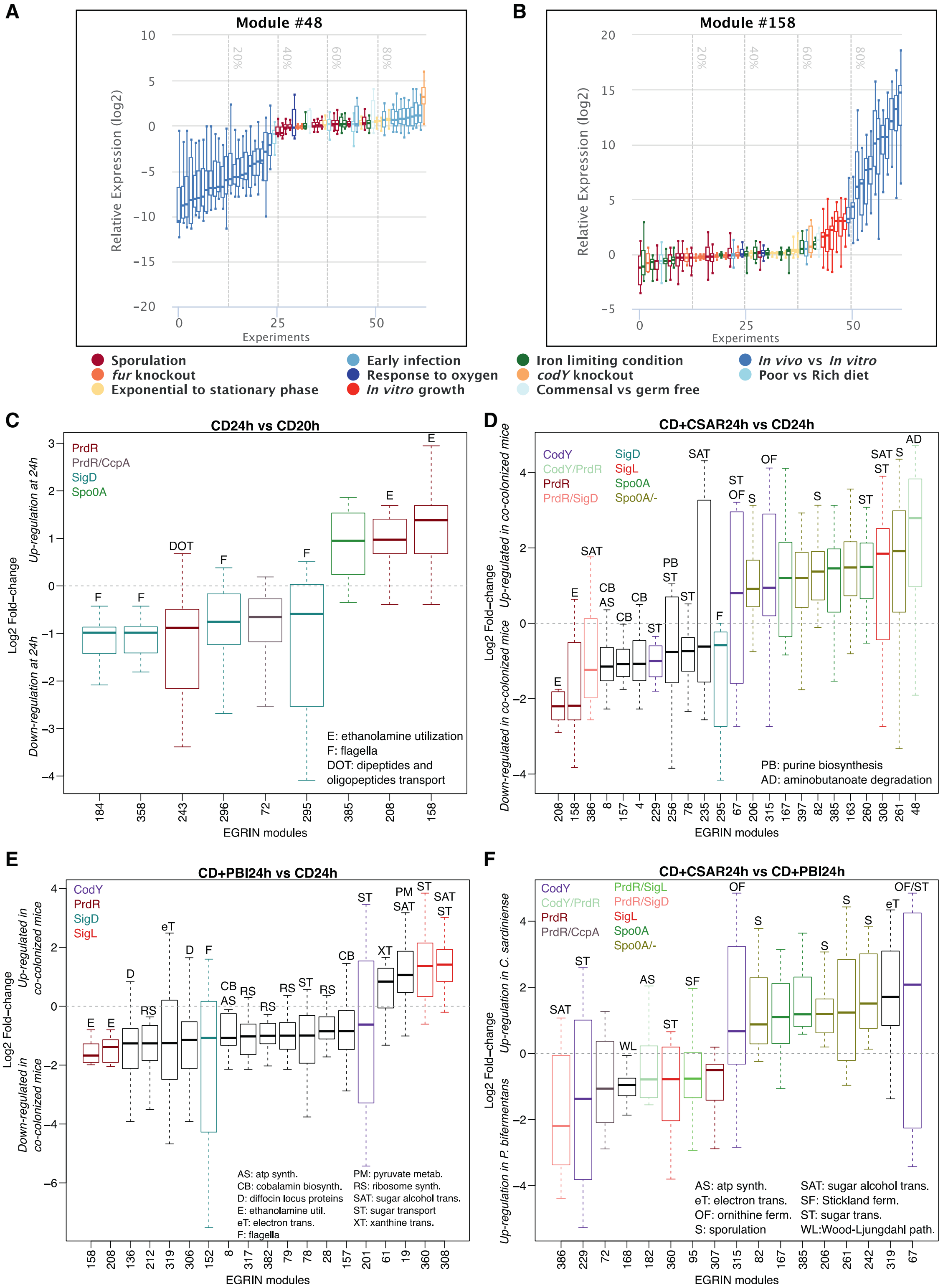


Figure 4. The EGRIN model identifies TFs driving the *in vivo* response of *C. difficile* when interacting with gut commensals

(A) Expression profile of module #48 across the transcriptional compendium used to build EGRIN. Each box represents log_2 fold-change (with respect to a dataset-specific reference, Table S1) of members of the module in a single transcriptome. Transcriptomes are color coded according to their membership in the 11 condition blocks in the compendium and ranked based on their median log_2 fold-change. The "early infection" condition block is statistically overrepresented in the 20% of highest (indicated with the 80% dashed line) transcriptomes median log_2 fold-change.

(legend continued on next page)

that drive *C. difficile*'s colonization and adaption to *in vivo* environments (Janoir et al., 2013; Janvilisri et al., 2010). Analyses identified 680 genes across 43 modules that were significantly upregulated *in vivo*, while 1,325 genes across 82 modules were significantly downregulated (STAR Methods; Data S1). Notably, module #48 (described above) was upregulated during early infection (Figure 4A). In contrast, module #158 was upregulated at later stages of infection (Figure 4B). This module is enriched for putative PrdR and EutV co-regulated ethanolamine utilization genes and contains *eut* operons that encode ethanolamine fermentative enzymes (Nawrocki et al., 2018) (Figure S3D). Ethanolamine is prevalent within gut secretions and is also released from damaged host tissues, providing a carbon and nitrogen source for *C. difficile*. The predicted co-regulation of this gene module by PrdR (Pearson correlation = -0.29 and p value = $9.7e-04$) suggests additional *in vivo* functions of PrdR to optimize *C. difficile* metabolism in gut environments. The negative relation between *prdR* expression and module #158 expression indicates that enhanced utilization of ethanolamine *in vivo* may contribute to the decreased survival observed in hamsters infected with a *prdR* deletion strain (Bouillaut et al., 2019).

With capacity to identify intestinal contributions to *C. difficile* responses, we leveraged the EGRIN model to analyze commensal modulation of the pathogen's virulence, using transcriptomic datasets from gnotobiotic mice monocolonized with the mouse-infective strain *C. difficile* ATCC43255 or co-colonized with *C. difficile* and the protective gut commensal species *Paraclostridium bifermentans*, or infection-worsening species *Clostridium sardiniense* (Girinathan et al., 2021). These datasets were not used in model construction. By mapping sets of differentially expressed genes into the EGRIN model we uncovered differential regulation of modules across 18 cellular processes and their associated TFs in the presence of *P. bifermentans* or *C. sardiniense* (Figures 4C–4F).

One Spo0A-enriched module (module #385) was upregulated by 24 h of infection in monocolonized mice (Figure 4C). The same module was upregulated by 24 h of infection in *C. sardiniense* co-colonized mice, in addition to three other modules enriched with sporulation genes and the Spo0A regulon (modules #82, #206, and #261 in Figure 4D). On the other hand, no sporulation-enriched modules were detected by 24 h of infection in *P. bifermentans* co-colonized mice (Figure 4E). Comparison of *C. sardiniense* co-colonized mice and *P. bifermentans* co-colonized mice discovered four sporulation-enriched modules (modules #82, #206, #242, and #261) in the virulent context with *C. sardiniense* (Figure 4F), a finding confirmed by higher spore biomass of *C. difficile* when co-colonized with *C. sardiniense* (Girinathan et al., 2021). This analysis suggested that the sporulation pathway is an indicator of *C. difficile* disease, reinforcing the Spo0A-mediated link between sporulation and toxin production recapitulated by the model (Figure 2F).

Module #319 contains genes associated with electron transport via Rnf ferredoxin systems, and steps in glycolytic, butanoate, and succinate metabolism (Figure S3E). This module was consistently downregulated in *P. bifermentans* co-colonized mice (Figure 4E), while it was upregulated in *C. sardiniense* co-colonized mice (Figure 4F). These findings highlight activation of multiple co-regulated energy-generating pathways in hypervirulent states of *C. difficile* infection, which may contribute to virulence and the pathogen's responses to changing redox states from host inflammatory responses. Remarkably, the EGRIN model identified the NAD^+/NADH sensing regulator Rex as a potential repressor of module #319 (Inferelator regression coefficient = -0.075). Hence, the observed downregulation of module #319 *in P. bifermentans* co-colonized mice may indicate increased Rex activity in *C. difficile* from an associated nutrient depleted state less able to support its growth (Girinathan et al., 2021).

Four modules enriched with SigD target genes encoding flagellar components (modules #184, #295, #296, and #358) were downregulated in monocolonized mice at 24 h (Figure 4C), indicating repression of motility to divert cellular resources toward pathogenesis. This finding is supported by increased virulence of *C. difficile* strains lacking a functional flagellum (Dingle et al., 2011). In summary, EGRIN analyses generated mechanistic insights and hypotheses for the interplay of specific genes and TFs for sporulation, energy production, and flagella synthesis that might underlie the enhanced and subdued virulence of *C. difficile* in different contexts.

Metabolic network analyses elucidate *in vivo* metabolic adaptations of *C. difficile*

To investigate how specific genes within *C. difficile* contribute to *in vivo* phenotypes needed to develop symptomatic infection, we extended a previously developed icdf834 metabolic model for CD630 (Kashaf et al., 2017; Larocque et al., 2014). We added 4 genes for molybdenum utilization and cofactor synthesis and exchange reactions to account for *C. difficile* capacity to utilize mannitol, fructose, sorbitol, raffinose, succinate, and butanoate (Janoir et al., 2013; Theriot et al., 2014). We refer to this updated model as icdf836 (Figure 5A). The model derived for CD630 was projected onto homologous genes and associated reactions in *C. difficile* ATCC43255. In total, 766 out of 836 genes (92%) in the icdf836 model were conserved across the two strains (Data S2).

We validated the completeness and accuracy of this model by confirming its ability to predict biomass production in three

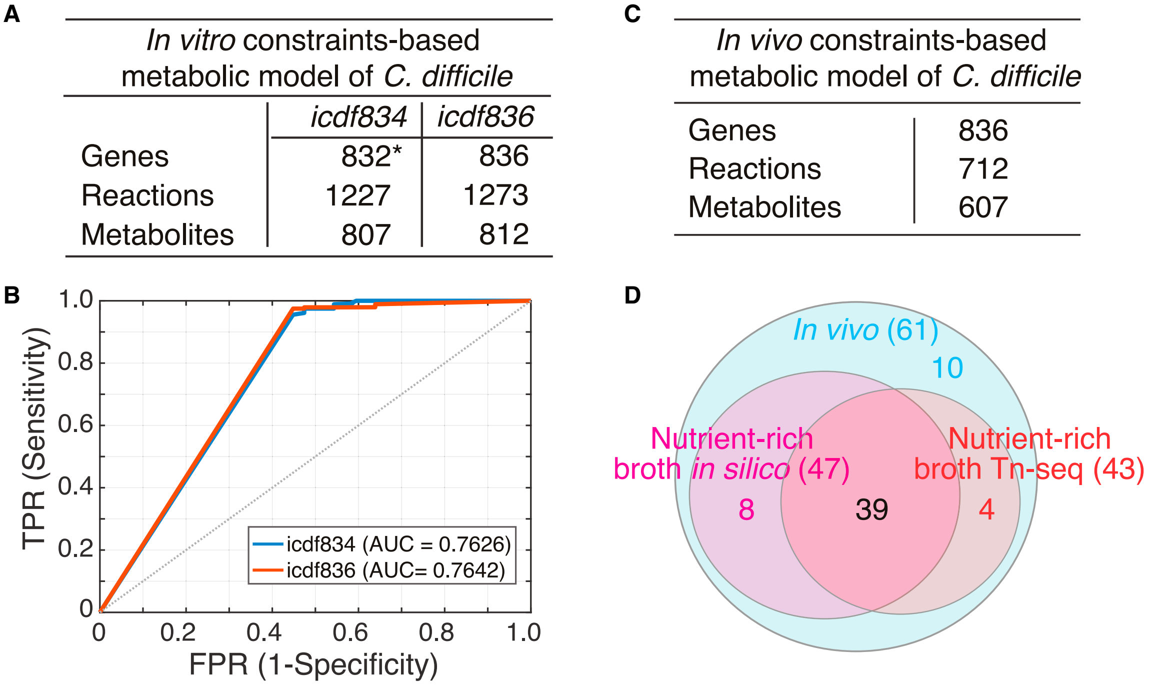
(B) Expression profile of module #158. Only one condition block ("*in vivo* versus *in vitro*") was found in the 20% of highest fold-change.

(C) EGRIN modules enriched with genes differentially expressed (absolute $\log_2$ fold-change > 1 and adjusted p value < 0.05) in *C. difficile* monocolonized mice at 24 versus 20 h of infection. x axis shows module IDs. Modules were annotated according to their functional enrichment and overlap with manually curated TF regulons (Table S2).

(D) Enriched EGRIN modules in *C. sardiniense+C. difficile* co-colonized mice versus *C. difficile* monocolonized mice at 24 h of infection. Due to space constraint, only abbreviations of functional terms not shown in other panels are displayed. Modules enriched with Spo0A, and sporulation sigma factors are indicated with "Spo0A/−."

(E) Enriched EGRIN modules in *P. bifermentans+C. difficile* co-colonized mice versus *C. difficile* monocolonized mice at 24-h of infection.

(F) Enriched EGRIN modules in *P. bifermentans+C. difficile* co-colonized mice versus *C. sardiniense+C. difficile* co-colonized mice at 24 h of infection. For all comparisons, only modules with absolute median fold-changes ≥ 0.5 and enriched with TF regulons or functional categories are displayed. In all panels, boxes cover the 25th-75th percentile range (with medians shown as horizontal lines).



In vitro constraints-based metabolic model of *C. difficile*

	icdf834	*icdf836*
Genes	832*	836
Reactions	1227	1273
Metabolites	807	812

In vivo constraints-based metabolic model of *C. difficile*

Genes	836
Reactions	712
Metabolites	607

Figure 5. Metabolic model predictions

(A) Details of the *in vitro* metabolic models (icdf834 and icdf836) of *C. difficile* 630 (Kashaf et al., 2017). General properties of the icdf836 model (generated in this study) after adding the required *in vivo* exchanges, transports and reactions are shown. The "*" indicates that there were two duplicated genes in the icdf834 model, reducing the total number of genes to 832.

(B) ROC curves showing the accuracy of icdf834- and icdf836-predicted gene essentiality in nutrient-rich medium evaluated against a Tn-seq functional screen (Dembek et al., 2015).

(C) Details of the *in vivo* (monocolonized) model derived using the GIMME algorithm (Becker and Palsson, 2008) on the icdf836 model where only the active reactions are included from *in vivo* transcriptome.

(D) Venn diagram showing the number of model-predicted essential genes for growth of *C. difficile* 630 *in vitro* versus *in vivo*. Only genes predicted as essential for *in vivo* (monocolonized) growth were considered in the analysis.

different growth media compositions: (1) minimal medium, (2) basal defined medium, and (3) complex, nutrient-rich medium (STAR Methods). The model accurately predicted *C. difficile*'s requirements for six amino acids: cysteine, leucine, isoleucine, proline, tryptophan, and valine (Karasawa et al., 1995). Using a Tn-seq generated genome-wide fitness screen (Dembek et al., 2015), we tested the accuracy of model-predicted importance of each gene in supporting the growth of *C. difficile* in nutrient-rich conditions using a receiver-operator characteristic (ROC) curve. The area under the ROC curves for the icdf834 and icdf836 models were 0.763 (p value = 0.015) and 0.764 (p value = 0.029), respectively (Figure 5B), indicating that both models significantly outperformed a random model.

We next extended and applied the model to predict *C. difficile* behaviors and gene essentiality *in vivo*. *C. difficile* transcriptomes from specifically colonized gnotobiotic mice (Girinathan et al., 2021) were input into the GIMME algorithm (Becker and Palsson, 2008). Per the *in vivo* gene expression, we determined that 712 of the 1,273 reactions in the icdf836 model were active and likely to be important for colonization and growth over the course of monocolonized infection (Figure 5C). The model made two notable predictions *in vivo* regarding the pathogen's metabolism. First, the icdf836 model predicted 14 amino acids to be required for *C. difficile* growth *in vivo*, in contrast to the 6 required *in vitro* (Data S2). These amino acids included the dominant Stickland-fermented amino acids that were also required *in vitro* (e.g., proline and branched-chain amino acids) and additional amino acids such as arginine, glutamate, lysine, and methionine, which also function in cell wall synthesis, nitrogen cycling, and responses to oxidative stress. Second, the model-predicted *C. difficile*'s switch from preferential use of glucose *in vitro* in complex media, to simultaneous utilization *in vivo* of diverse carbohydrate sources including fructose, galactose, maltose, and sugar alcohols (e.g., mannitol and sorbitol) to promote colonization and growth (Data S2). Seven of these carbohydrate sources were described in other mouse infection studies illustrating support for these findings across *C. difficile* strains, and in germ-free and conventional mouse models (Data S2) (Janoir et al., 2013; Jenior et al., 2017; Theriot et al., 2014).

Integration of EGRIN and metabolic model reveals the interplay of regulation and metabolism during *C. difficile in vivo* adaptations

To evaluate how transcriptional regulation modulates *C. difficile* metabolic and physiological responses, we integrated the transcriptional and metabolic networks into a PRIME model. PRIME evaluates context-specific TF essentiality (Immanuel et al., 2021).

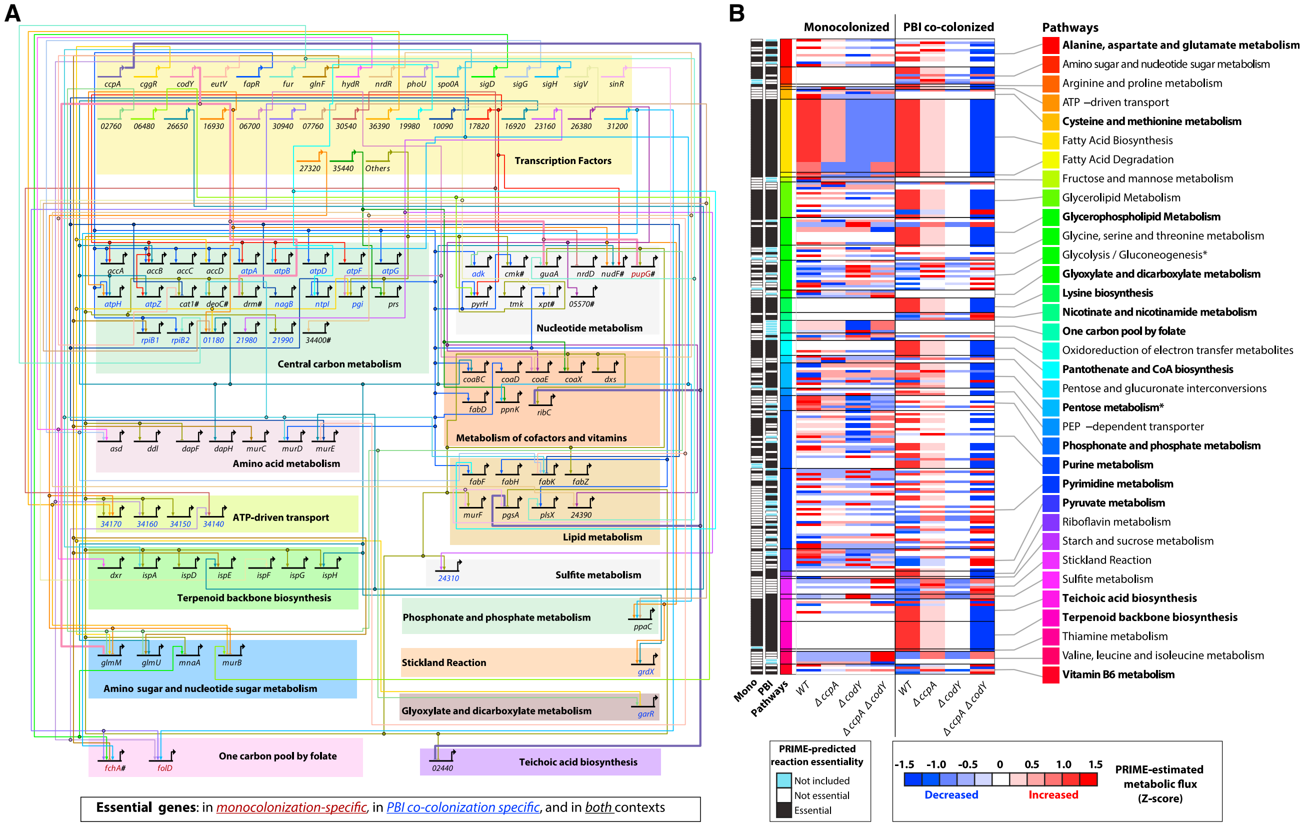

A
ccpA
cggR
codY
eutV
fapR
fur
glnF
hydR
nrdR
phoU
spo0A
sigD
sigG
sigH
sigV
sinR
02760
06480
26650
16930
06700
30940
07760
30540
36390
19980
10090
17820
16920
23160
26380
31200
27320
35440
Others
Transcription Factors
accA
accB
accC
accD
atpA
atpB
atpD
atpF
atpG
atpH
atpZ
cat1#
deoC#
drm#
nagB
ntpI
pgi
prs
rpiB1
rpiB2
01180
21980
21990
34400#
Central carbon metabolism
adk
cmk#
guaA
nrdD
nudF#
pupG#
pyrH
tmk
xpt#
05570#
Nucleotide metabolism
coaBC
coaD
coaE
coaX
dxs
fabD
ppnK
ribC
Metabolism of cofactors and vitamins
asd
ddl
dapF
dapH
murC
murD
murE
Amino acid metabolism
fabF
fabH
fabK
fabZ
murF
pgsA
plsX
24390
Lipid metabolism
34170
34160
34150
34140
ATP-driven transport
24310
Sulfite metabolism
dxr
ispA
ispD
ispE
ispF
ispG
ispH
Terpenoid backbone biosynthesis
Phosphonate and phosphate metabolism
ppaC
Stickland Reaction
grdX
glmM
glmU
mnaA
murB
Amino sugar and nucleotide sugar metabolism
Glyoxylate and dicarboxylate metabolism
garR
fchA#
folD
One carbon pool by folate
02440
Teichoic acid biosynthesis
Essential genes: in monocolonization-specific, in PBI co-colonization specific, and in both contexts
B
Monocolonized
PBI co-colonized
Pathways
Alanine, aspartate and glutamate metabolism
Amino sugar and nucleotide sugar metabolism
Arginine and proline metabolism
ATP –driven transport
Cysteine and methionine metabolism
Fatty Acid Biosynthesis
Fatty Acid Degradation
Fructose and mannose metabolism
Glycerolipid Metabolism
Glycerophospholipid Metabolism
Glycine, serine and threonine metabolism
Glycolysis / Gluconeogenesis*
Glyoxylate and dicarboxylate metabolism
Lysine biosynthesis
Nicotinate and nicotinamide metabolism
One carbon pool by folate
Oxidoreduction of electron transfer metabolites
Pantothenate and CoA biosynthesis
Pentose and glucuronate interconversions
Pentose metabolism*
PEP –dependent transporter
Phosphonate and phosphate metabolism
Purine metabolism
Pyrimidine metabolism
Pyruvate metabolism
Riboflavin metabolism
Starch and sucrose metabolism
Stickland Reaction
Sulfite metabolism
Teichoic acid biosynthesis
Terpenoid backbone biosynthesis
Thiamine metabolism
Valine, leucine and isoleucine metabolism
Vitamin B6 metabolism
Mono
PBI
Pathways
WT
Δ ccpA
Δ codY
Δ ccpA Δ codY
WT
Δ ccpA
Δ codY
Δ ccpA Δ codY
PRIME-predicted reaction essentiality
Not included
Not essential
Essential
-1.5
-1.0
-0.5
0
0.5
1.0
1.5
Decreased
Increased
PRIME-estimated metabolic flux (Z-score)

Figure 6. PRIME-estimated metabolic fluxes and gene essentiality in mono- and *P. bifermentans* co-colonized conditions

(A) BioTapestry visualization of *in vivo* gene regulatory network for *C. difficile* 630: 79 essential genes with Inferelator-predicted transcriptional regulators are shown. The genes and regulators shown as five-digit numbers represent the nomenclature preceded by "CD630_." All regulators with less than two essential targets were combined in the "others" meta-regulator. The "#" symbol indicates the 10 genes predicted as essential in the monocolonized condition but not essential *in vitro*.

(B) PRIME-estimated metabolic fluxes for each reaction in the mono- and co-colonized conditions. *Z* score transformation was independently applied to each reaction and condition. The "*" symbol indicates that there is some contention about the existence of this pathway in *C. difficile*, and it may need to be revised in future models.

We first inferred a transcriptional network of 1,401 TF-metabolic gene interactions, using the transcriptome compendium compiled for the CD630 strain. This step inferred combinatorial and weighted regulatory influences of 215 TFs on 731 metabolic genes as input for PRIME to model downstream consequences on metabolic flux through associated enzymatic reactions. We used the CD630 models to generate orthologous models for ATCC43255 by leveraging the high degree of conservation of TFs (97%) and metabolic genes (92%) between the two strains. Finally, we generated condition-specific PRIME models for ATCC43255 using transcriptomic data (for genes conserved in both strains) from mono- and *P. bifermentans* co-colonized mice. This approach uncovered how CcpA and CodY act synergistically during *in vivo* adaptation of ATCC43255, especially in the presence of *P. bifermentans*. Thus, the model projection strategy is useful to study evolutionary conserved adaptive mechanisms across strains of *C. difficile*. However, this strategy will not be useful to study strain-specific phenotypes, which would require strain-specific network models. Comparative analysis of datasets and network models across strains will be ultimately needed to elucidate how genomic differences across strains manifest in distinct clinical phenotypes and outcomes.

PRIME predicted conditionally essential metabolic genes and networks that promote *C. difficile*'s growth *in vivo* (i.e., in monocolonized mice). In simulating the consequences of single gene deletions, PRIME predicted that pathogen growth in monocolonized mouse would decrease by $\geq$ 65% with individual knockouts of 10 genes involved in one-carbon cycling reactions, and nucleotide and central carbohydrate metabolism (Figures 5D and 6A; Table S5). These metabolic pathways represent potential targets that drive *C. difficile*'s colonization and other factors required to develop symptomatic infections. PRIME also predicted *C. difficile*'s shift from carbohydrate utilization toward amino acid utilizing pathways *in vivo*, per the enhanced set of 14 amino acids, including the preferred Stickland-related amino acids (leucine and proline) known to support metabolism and growth (Janoir et al., 2013; Jenior et al., 2017; Theriot et al., 2014). Many of these amino acids show high abundance within the gut lumen in gnotobiotic and in antibiotic-treated conventional mice that enhance *C. difficile* capacity to colonize and expand (Girinathan et al., 2021). Using transcriptome profiles of *C. difficile* from *P. bifermentans* co-colonized mice, PRIME predicted that *P. bifermentans* inhibits *C. difficile* biomass at least 3.6-fold, in agreement with the 3.2-fold inhibition observed

Table 1. Experimental validation of PRIME-predicted synergistic epistasis between CcpA and CodY

	Monocolonized				PBI co-colonized			
	PRIME prediction		Experiment		PRIME prediction		Experiment	
	Relative growth		Relative growth (Log$_{10}$)[a,c]		Relative growth		Relative growth (Log$_{10}$)[a,c]	
TF KO strain	(Log$_{10}$)[a]	Essentiality[b]	24 h	Essentiality[b]	(Log$_{10}$)[a]	Essentiality[b]	24 h	Essentiality[b]
Δ*ccpA*	−0.716	essential	0.609*	non-essential	−0.144	non-essential	0.315#	non-essential
Δ*codY*	−0.482	essential	−1.181**	essential	−0.176	non-essential	0.527#	non-essential
Δ*ccpA* Δ*codY*	−1.198	essential	−1.299***	essential	−0.885	Essential	−1.708***	essential

[a] Relative growth with respect to wild-type genotype.

[b] Gene deletions that reduce *C. difficile* growth by more than 65% (i.e., $\log_{10}$(relative growth) < -0.456) were labeled as essential.

[c] T test p values $\geq$ 0.05 are indicated with the "#" symbol and were considered not significant. The "*," "**," and "***" symbols indicate p values < 0.05, < 0.01, and < 0.001, respectively.

experimentally (Girinathan et al., 2021). PRIME also predicted that the growth inhibition is likely a consequence of *P. bifermentans* scavenging essential nutrients, including glucose and Stickland-fermentable amino acids, which drives metabolic remodeling in the pathogen by reducing flux through reactions required *in vivo* for biomass production (Figure 6; Data S2).

A key advantage of PRIME is its extended capability to evaluate TF essentiality. We evaluated PRIME performance by generating ROC curves for gene essentiality predictions. The area under the ROC curve was comparable for the icdf834 and icdf836 models (Figure S4). Given CcpA and CodY influence on toxin production and metabolic genes (Antunes et al., 2012; Dineen et al., 2010), we leveraged PRIME to predict the phenotypic consequences of single and double knockouts in *ccpA* and *codY* in mono- and *P. bifermentans* co-colonized mice (Table 1). PRIME predicted that three genes regulated by CcpA (*pgsA*, *ribC*, and CD630_02440) were essential in the mono- and co-colonized conditions (Figure 6A). Similarly, PRIME predicted that three genes regulated by CodY (*drm*, *glmM*, *guaA*) were essential in mono- and co-colonized mice. One additional gene (*pupG*) was only essential in monocolonized mice (Figure 6A). Strikingly, PRIME predicted the double knockout of CcpA and CodY to synergistically inhibit *C. difficile* growth (as compared with single TF knockouts), especially in *P. bifermentans* co-colonized mice. Five of the six PRIME essentiality predictions were validated in germ-free and *P. bifermentans*-colonized mice infected with wild-type or mutant strains of *C. difficile* (Table 1; Figure S5). PRIME also offered mechanistic insights into the conditional phenotypes of *C. difficile* by revealing how the single and double knockouts of *ccpA* and *codY* remodeled the metabolic state of the pathogen in mono- and co-colonized contexts (Figure 6B). For example, reactions associated with the one-carbon pool by the folate pathway were inactive in the co-colonized condition but were differentially affected by the gene knockouts in the monocolonized mouse. PRIME also predicted that flux for reactions associated with teichoic acid biosynthesis and terpenoid backbone biosynthesis was invariable in the monocolonized mice but sensitive to gene deletions in co-colonized mice. In summary, by delineating how transcriptional regulation propagates throughout the metabolic network to manifest in fitness, PRIME offers the capability to investigate how *C. difficile* adapts to complex biotic and abiotic changes within the *in vivo* environment.

The *C. difficile* web portal, a resource for the *C. difficile* community

We have released the *C. difficile* web portal (http://networks.systemsbiology.net/cdiff-portal/) to provide a discovery and collaboration gateway for the *C. difficile* scientific community. The portal aims to accelerate the advancement of the science and understanding of *C. difficile* biology, and the relationships among gene regulation, metabolism, and virulence. Within the portal, users can access publicly available datasets (e.g., transcriptional compendia), models, software, and supporting resources. The portal includes information on more than 4,000 *C. difficile* genes, 1,273 metabolic reactions, and 406 EGRIN modules. The EGRIN model can be interactively explored and queried with gene set enrichment analysis to elucidate regulatory networks that may be relevant for a user-specified group of genes. Genes can be explored in the context of genome annotations, expression profiles, regulatory and metabolic membership, and other functional genomic information across databases including COG, Uniprot, and PATRIC (Consortium, 2017; Galperin et al., 2015; Wattam et al., 2014). The portal provides access to detailed information on (1) genes, (2) predicted gene modules, and (3) metabolic reactions (Figure S6**A**).

Each module page includes summary statistics for the module, context-specific differential activity patterns, GREs, TF regulatory influences, enrichment of biological functions and pathways, and information on each module member gene. The module pages are structured to facilitate the assessment of the quality and statistical significance of the modules and highlight functional connections, while allowing users to implement their own filters (e.g., regression coefficient and adjusted p value thresholds) (Figure S6**B**). The portal includes a table of metabolic reactions with details of each reaction, associated genes, metabolites, and sub-systems. Metabolites and subsystems are defined as taxonomic vocabularies that collect and group associated reactions to identify related metabolic processes. In addition, the portal provides access to algorithms, software, and data and will include information about animal models, strains, and other *C. difficile*-relevant community resources. As additional datasets are communicated, model predictions and tools will be successively updated to support systems-level analyses and assist in hypothesis generation in *C. difficile* biology and to enable tangible clinical interventions.

DISCUSSION

C. difficile is unique among gut anaerobes in possessing a diverse carbon source metabolism to enable colonization and growth in gut environments. These systems further exist within a complex network of gene regulatory modules that modulate growth, energy balance, and stress responses. Capacity to understand these systems-level integration points has remained challenging in the absence of robust systems biology models to infer *C. difficile*'s *in vivo* behaviors. We acknowledge the detailed studies from multiple groups over prior decades that provided a critical mass of information on *C. difficile*'s nutrient and gene-level responses to support development of an EGRIN model for a gut anaerobe and toxigenic species. We emphasize that this information, the most for any obligate anaerobe, still represents a small fraction of that normally used to develop thorough EGRIN models. Despite this caveat, the EGRIN model was effective in uncovering functional and mechanistic insights into regulatory and metabolic responses of the pathogen in contexts not covered by the training transcriptome dataset—both *in vitro* and *in vivo* (i.e., in the presence of commensals). While this demonstrates the utility of the EGRIN model to analyze new transcriptomes that were not used for training (Figures 4C–4F and S7), it also underscores the potential for further improving the model with new data and experimental validations. In the future, we will improve model coverage and performance using an ensemble modeling approach with a larger compendium of transcriptomes from new infection- and treatment-relevant contexts to identify condition-specific regulatory networks with probabilistic association across genes, environmental contexts, and regulatory mechanisms (Brooks et al., 2014). Similarly, performance for the icdf834 model (extended into the icdf836 model in this study) in predicting gene essentiality in an *in vitro* context was similar to that of a new metabolic model for *C. difficile* (iCN900, Norsigian et al. 2020), published during the development of the PRIME model, with area under the ROC curve of 0.76 and 0.7, respectively. We will periodically update both EGRIN and PRIME models to incorporate new transcriptomes and improved metabolic models. The models will also be refined using new tools for genetically manipulating *C. difficile* strains, including ATCC43255 (Peltier et al., 2020), which will enable targeted validations of regulatory influences of TFs and GREs on critical aspects of its metabolism, growth, and virulence.

The *C. difficile* EGRIN model enables a number of predictions relevant to *in vivo* disease. For example, PrdR, a regulator of the Stickland proline reductase (*prdABDEF*) and other genes, has long been hypothesized to have a role in PaLoc gene expression through as-yet unknown mechanisms. EGRIN identified combined PrdR and CodY effects on *tcdA* gene expression (suggested by the enrichment of module #182 with both PrdR and CodY regulons), providing a regulatory integration point and broader set of co-regulated genes to support further experiments. Biclustering also identified interactions between Spo0A, another regulator hypothesized to modulate PaLoc expression, and *tcdB* expression in module #397. The identified modules, associated genes and regulators provide information to support further experimental investigation of combinatorial effects of these and other regulators identified in PaLoc gene-associated modules. The EGRIN model also predicted PrdR has an important role *in vivo* based on its systems-level effects on critical metabolic and regulatory networks supporting colonization, metabolism, and growth. These PrdR-mediated interactions involve multiple direct and indirect effects upon other modules (e.g., module #48 and module #158) and aspects of the pathogen's metabolism and gene regulation.

The current model did not identify all previously characterized regulators of PaLoc expression, including SigD regulation of TcdR, and effects of other more recently identified PaLoc regulators such as RstA and LexA, for which limited datasets exist from wild-type or mutant strains cultured under relevant nutritional and other environmental contexts. Additional datasets with isogenic regulator mutants will likely improve predictions while further validating previously defined biologic effects. Nonetheless, as shown with our *in vivo* analyses, application of the EGRIN and PRIME models to new datasets offers key insights into causal mechanistic drivers of adaptive strategies of the pathogen. Notably, the CD630 model was trained on <10% of transcriptome information and <2% of ChIP-seq datasets used for constructing EGRIN models for other organisms. Yet, both models demonstrated high levels of accuracy in recapitulating previously characterized regulatory and metabolic phenomena associated with *C. difficile* growth *in vitro* and mouse infection and colonization. Thus, the predictive capabilities of EGRIN and PRIME serve as formative tools to uncover biological insights through hypothesis-driven design of experiments to characterize regulatory and metabolic mechanisms that are essential for infection and colonization by *C. difficile* (e.g., CD630_16930 and CD630_17820 are putative regulators of PRIME-predicted essential genes; Figure 6A). This model-driven design of experiments will iteratively improve the coverage and accuracy of regulatory and metabolic mechanisms modeled by EGRIN and PRIME. We can also improve the models significantly by expanding the datasets with experiments that probe transcriptional and metabolic responses of the pathogen in infection-, host-, and treatment-relevant contexts that are poorly represented in the current compendium of transcriptomes. Some of the poorly represented conditions in the transcriptome compendium include *in vivo* responses of *C. difficile* to antibiotic treatment, as well as its responses to host-relevant conditions such as oxidative stress, acidic pH, and nutrient starvation (Edwards et al., 2016b). Thus, EGRIN and PRIME will serve as a community-wide resource for model-driven experimentation that will iteratively advance systems-level understanding of adaptive strategies employed by *C. difficile* to infect and colonize the host.

Leveraging additional Tn-seq and *in vivo* transcriptomic datasets, the expanded icdf836 model identified a broader set of amino acids, in addition to genes and anaerobe-specific pathways, needed to support colonization and growth expansion *in vivo*. Notably, predictions of *in vivo* gene essentially identified folate one-carbon cycling pathways including those connected with Wood-Ljundahl fixation of carbon dioxide to acetate (Gössner et al., 2008). Predictions of gene essentiality also identified multiple nucleotide synthesis and salvage pathway genes that were essential *in vivo* but not *in vitro*, including ones associated with xanthine transport and metabolism, an abundant nucleotide in gut secretions that originates from host sources (Girinathan et al., 2021). Lastly, the systems approach using the two models has identified genes and TFs across disparate pathways, including sporulation, flagella biosynthesis, sugar metabolism, and biosynthesis of aromatic and branched-chain amino acids, which contribute to growth in an intestinal environment. Once experimentally validated, these essential genes represent vulnerabilities that can be rationally targeted with small molecules, bacteriotherapeutics, or other patient interventions. We believe that by democratizing the disparate data, algorithms, and models through interactive exploration capabilities, the *C. difficile* web portal will accelerate collaborative systems analysis of host-pathogen-commensal interactions by engaging the wider scientific community.

We illustrate additional predictions from the *C. difficile* EGRIN model to enable gene- through systems-level analyses of the pathogen. The *C. difficile* genome still contains a high number of genes of unknown function. Model predictions allowed us to assign putative functions to 48 genes, including ones associated with sporulation, carbohydrate transport, and metabolism. Integration of the regulatory and metabolic network models with PRIME enabled unprecedented insight into the essential role of TFs in mediating combinatorial control of metabolism in different colonization contexts, vis-à-vis presence and absence of the protective commensal *P. bifermentans*. For instance, PRIME accurately predicted synergistic epistasis between the CodY and CcpA networks, a phenotype that was not attributable to a single downstream metabolic gene, reaction, or process. Rather, the significantly diminished growth of a double knockout of CodY and CcpA (especially in *P. bifermentans* co-colonized mice) appears to emerge from the interplay of more than 35 TFs regulating ~80 genes catalyzing reactions across more than two dozen metabolic processes. Furthermore, by uncovering conditionally essential genes and pathways, PRIME can be used to understand how *C. difficile* escapes therapies and develop strategies to block these escape routes with synergistically acting secondary antibiotics. The *C. difficile* web portal makes all of the tools and resources available to the broader research community, providing a platform for collaboration and to support systems-level investigations of the pathogen and its interactions with the host and commensal microbiota.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
  - ○ *C. difficile* genome annotation
  - ○ *C. difficile* transcriptional compendium
  - ○ Construction of the EGRIN model
  - ○ Literature-derived TF regulons
  - ○ DNA motif comparison
  - ○ Function assignment to uncharacterized genes
  - ○ Analysis of *in vivo* data
  - ○ Analysis of transcriptomes not used for training
  - ○ Metabolic model refinement
  - ○ Gene essentiality prediction
  - ○ PRIME model development
  - ○ Network visualization
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Module enrichment evaluation
  - ○ Transcriptional profiles of EGRIN modules
- ADDITIONAL RESOURCES
  - ○ The *C. difficile* web portal

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.chom.2021.09.008.

ACKNOWLEDGMENTS

We thank members of the Baliga lab for critical discussions and feedback. Funding was provided by National Institute of Allergy and Infectious Diseases (R01AI128215, R01AI141953, and U19AI135976), and the National Science Foundation (DBI-1565166 and IIBR-2042948) to N.S.B., and the BWH Precision Medicine Institute, R01AI153605 from the National Institute of Allergy and Infectious Diseases (NIAID) and P30DK034854 from the National Institute of Diabetes, Digestive and Kidney Diseases to L.B. J.N.W. was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. Funding was also provided by the Institut Pasteur and Université de Paris to B.D. and J.P., and the Institut Universitaire de France to O.S.

AUTHOR CONTRIBUTIONS

Conceptualization, M.L.A.-O., L.B., and N.S.B.; data curation, M.L.A.-O., S.R.C.I., S.T., B.P.G., J.N.W., N.D., J.P., O.S., B.D., and L.B.; formal analysis, M.L.A.-O. and S.R.C.I.; investigation, M.L.A.-O., S.R.C.I., J.P., B.D., L.B., and N.S.B.; methodology, M.L.A.-O., S.R.C.I., S.T., and W.-J.W.; project administration, M.L.A.-O. and N.S.B.; resources, J.P., B.D., L.B., and N.S.B.; software, M.L.A.-O., S.T., and W.-J.W.; supervision, N.S.B.; visualization, M.L.A.-O., S.R.C.I., S.T., W.-J.W., and N.S.B.; writing, M.L.A.-O., S.R.C.I., S.T., L.B., and N.S.B.

DECLARATION OF INTERESTS

L.B. is an inventor on patents for *C. difficile* microbiota therapeutics. L.B. is an SAB member and holds stock in ParetoBio. Remaining authors declare no competing interests.

Received: May 18, 2021
Revised: July 29, 2021
Accepted: September 16, 2021
Published: October 11, 2021

REFERENCES

Aktories, K. (2011). Bacterial protein toxins that modify host regulatory GTPases. Nat. Rev. Microbiol. *9*, 487–498.

Antunes, A., Camiade, E., Monot, M., Courtois, E., Barbut, F., Sernova, N.V., Rodionov, D.A., Martin-Verstraete, I., and Dupuy, B. (2012). Global transcriptional control by glucose and carbon regulator CcpA in Clostridium difficile. Nucleic Acids Res *40*, 10701–10718.

Antunes, A., Martin-Verstraete, I., and Dupuy, B. (2011). CcpA-mediated repression of Clostridium difficile toxin gene expression. Mol. Microbiol. *79*, 882–899.

Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D.S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res *44*, W16–W21.

Arrieta-Ortiz, M.L., Hafemeister, C., Bate, A.R., Chu, T., Greenfield, A., Shuster, B., Barry, S.N., Gallitto, M., Liu, B., Kacmarczyk, T., et al. (2015).

An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. Mol. Syst. Biol. *11*, 839.

Arrieta-Ortiz, M.L., Hafemeister, C., Shuster, B., Baliga, N.S., Bonneau, R., and Eichenberger, P. (2020). Inference of bacterial small RNA regulatory networks and integration with transcription factor-driven regulatory networks. mSystems *5*, e00057.

Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME suite. Nucleic Acids Res *43*, W39–W49.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res *41*, D991–D995.

Becker, S.A., and Palsson, B.O. (2008). Context-specific metabolic networks are consistent with experiments. PLoS Comput. Biol. *4*, e1000082.

Berges, M., Michel, A.-M., Lassek, C., Nuss, A.M., Beckstette, M., Dersch, P., Riedel, K., Sievers, S., Becher, D., Otto, A., et al. (2018). Iron regulation in Clostridioides difficile. Front. Microbiol. *9*, 3183.

Boekhoud, I.M., Michel, A.-M., Corver, J., Jahn, D., and Smits, W.K. (2020). Redefining the Clostridioides difficile σB regulon: σB activates genes involved in detoxifying radicals that can result from the exposure to antimicrobials and hydrogen peroxide. mSphere *5*, e00728.

Bonneau, R., Facciotti, M.T., Reiss, D.J., Schmid, A.K., Pan, M., Kaur, A., Thorsson, V., Shannon, P., Johnson, M.H., Bare, J.C., et al. (2007). A predictive model for transcriptional control of physiology in a free living cell. Cell *131*, 1354–1365.

Bouillaut, L., Dubois, T., Francis, M.B., Daou, N., Monot, M., Sorg, J.A., Sonenshein, A.L., and Dupuy, B. (2019). Role of the global regulator Rex in control of NAD^+-regeneration in Clostridioides (Clostridium) difficile. Mol. Microbiol. *111*, 1671–1688.

Bouillaut, L., Self, W.T., and Sonenshein, A.L. (2013). Proline-dependent regulation of Clostridium difficile Stickland metabolism. J. Bacteriol. *195*, 844–854.

Bradshaw, W.J., Bruxelle, J.-F., Kovacs-Simon, A., Harmer, N.J., Janoir, C., Péchiné, S., Acharya, K.R., and Michell, S.L. (2019). Molecular features of lipoprotein CD0873: a potential vaccine against the human pathogen Clostridioides difficile. J. Biol. Chem. *294*, 15850–15861.

Bradshaw, W.J., Kirby, J.M., Roberts, A.K., Shone, C.C., and Acharya, K.R. (2017). The molecular structure of the glycoside hydrolase domain of Cwp19 from Clostridium difficile. FEBS Journal *284*, 4343–4357.

Brooks, A.N., Reiss, D.J., Allard, A., Wu, W.-J., Salvanha, D.M., Plaisier, C.L., Chandrasekaran, S., Pan, M., Kaur, A., and Baliga, N.S. (2014). A system-level model for the microbial regulatory genome. Mol. Syst. Biol. *10*, 740.

Brooks, A.N., Turkarslan, S., Beer, K.D., Lo, F.Y., and Baliga, N.S. (2011). Adaptation of cells to new environments. Wiley Interdiscip. Rev. Syst. Biol. Med. *3*, 544–561.

Dehal, P.S., Joachimiak, M.P., Price, M.N., Bates, J.T., Baumohl, J.K., Chivian, D., Friedland, G.D., Huang, K.H., Keller, K., Novichkov, P.S., et al. (2010). MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic Acids Res *38*, D396–D400.

Dembek, M., Barquist, L., Boinett, C.J., Cain, A.K., Mayho, M., Lawley, T.D., Fairweather, N.F., and Fagan, R.P. (2015). High-throughput analysis of gene essentiality and sporulation in Clostridium difficile. mBio *6*, e02383.

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

Dineen, S.S., Villapakkam, A.C., Nordman, J.T., and Sonenshein, A.L. (2007). Repression of Clostridium difficile toxin gene expression by CodY. Mol. Microbiol. *66*, 206–219.

Dingle, T.C., Mulvey, G.L., and Armstrong, G.D. (2011). Mutagenic analysis of the Clostridium difficile flagellar proteins, FliC and FliD, and their contribution to virulence in hamsters. Infect. Immun. *79*, 4061–4067.

Dubois, T., Dancer-Thibonnier, M., Monot, M., Hamiot, A., Bouillaut, L., Soutourina, O., Martin-Verstraete, I., and Dupuy, B. (2016). Control of Clostridium difficile physiopathology in response to cysteine availability. Infect. Immun. *84*, 2389–2405.

Edwards, A.N., Karim, S.T., Pascual, R.A., Jowhar, L.M., Anderson, S.E., and McBride, S.M. (2016b). Chemical and stress resistances of Clostridium difficile spores and vegetative cells. Front. Microbiol. *7*, 1698.

Edwards, A.N., Tamayo, R., and McBride, S.M. (2016a). A novel regulator controls Clostridium difficile sporulation, motility and toxin production. Mol. Microbiol. *100*, 954–971.

El Meouche, I., Peltier, J., Monot, M., Soutourina, O., Pestel-Caron, M., Dupuy, B., and Pons, J.-L. (2013). Characterization of the SigD regulon of C. difficile and its positive control of toxin production through the regulation of tcdR. PLoS One *8*, e83748.

Elena, S.F., and Lenski, R.E. (2003). Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat. Rev. Genet. *4*, 457–469.

Fimlaid, K.A., Bond, J.P., Schutz, K.C., Putnam, E.E., Leung, J.M., Lawley, T.D., and Shen, A. (2013). Global analysis of the sporulation pathway of Clostridium difficile. PLoS Genet *9*, e1003660.

Galperin, M.Y., Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res *43*, D261–D269.

Giordano, N., Hastie, J.L., and Carlson, P.E. (2018). Transcriptomic profiling of Clostridium difficile grown under microaerophillic conditions. Pathog. Dis. *76*, fty010.

Girinathan, B.P., DiBenedetto, N., Worley, J.N., Peltier, J., Arrieta-Ortiz, M., Immanuel, S.R.C., Lavin, R., Delaney, M.L., Cummins, C., Onderdonk, A.B., et al. (2021). The mechanisms of in vivo commensal control of *Clostridioides difficile* virulence. Cell Host Microbe *29*. Published online October 11. https://doi.org/10.1016/j.chom.2021.09.007.

Gössner, A.S., Picardal, F., Tanner, R.S., and Drake, H.L. (2008). Carbon metabolism of the moderately acid-tolerant acetogen Clostridium drakei isolated from peat. FEMS Microbiol. Lett. *287*, 236–242.

Govind, R., and Dupuy, B. (2012). Secretion of Clostridium difficile toxins A and B requires the holin-like protein TcdE. PLoS Pathog *8*, e1002727.

Hastie, J.L., Hanna, P.C., and Carlson, P.E. (2018). Transcriptional response of Clostridium difficile to low iron conditions. Pathog. Dis. *76*, fty009.

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S.N., Richelle, A., Heinken, A., Haraldsdóttir, H.S., Wachowiak, J., Keating, S.M., Vlasov, V., et al. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v. 3.0. Nat. Protoc. *14*, 639–702.

Ho, T.D., and Ellermeier, C.D. (2015). Ferric uptake regulator fur control of putative iron acquisition systems in Clostridium difficile. J. Bacteriol. *197*, 2930–2940.

Hofmann, J.D., Biedendieck, R., Michel, A.-M., Schomburg, D., Jahn, D., and Neumann-Schaal, M. (2021). Influence of L-lactate and low glucose concentrations on the metabolism and the toxin formation of Clostridioides difficile. PLoS One *16*, e0244988.

Hofmann, J.D., Otto, A., Berges, M., Biedendieck, R., Michel, A.-M., Becher, D., Jahn, D., and Neumann-Schaal, M. (2018). Metabolic reprogramming of Clostridioides difficile during the stationary phase with the induction of toxin production. Front. Microbiol. *9*, 1970.

Immanuel, S.R.C., Arrieta-Ortiz, M.L., Ruiz, R.A., Pan, M., de Lomana, A.L.G., Peterson, E.J.R., and Baliga, N.S. (2021). Quantitative prediction of conditional vulnerabilities in regulatory and metabolic networks of. Mycobacterium tuberculosis.bioRxiv. https://doi.org/10.1101/2021.01.29.428876.

Jackson, S., Calos, M., Myers, A., and Self, W.T. (2006). Analysis of proline reduction in the nosocomial pathogen Clostridium difficile. J. Bacteriol. *188*, 8487–8495.

Janoir, C., Denève, C., Bouttier, S., Barbut, F., Hoys, S., Caleechum, L., Chapetón-Montes, D., Pereira, F.C., Henriques, A.O., Collignon, A., et al. (2013). Adaptive strategies and pathogenesis of Clostridium difficile from in vivo transcriptomics. Infect. Immun. *81*, 3757–3769.

Janvilisri, T., Scaria, J., and Chang, Y.-F. (2010). Transcriptional profiling of Clostridium difficile and Caco-2 cells during infection. J. Infect. Dis. *202*, 282–290.

Jenior, M.L., Leslie, J.L., Young, V.B., and Schloss, P.D. (2017). Clostridium difficile colonizes alternative nutrient niches during infection across distinct murine gut microbiomes. mSystems *2*, e00063.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res *45*, D353–D361.

Karasawa, T., Ikoma, S., Yamakawa, K., and Nakamura, S. (1995). A defined growth medium for Clostridium difficile. Microbiology (Reading) *141*, 371–375.

Kashaf, S.S., Angione, C., and Lió, P. (2017). Making life difficult for Clostridium difficile: augmenting the pathogen's metabolic model with transcriptomic and codon usage data for better therapeutic target characterization. BMC Syst. Biol. *11*, 25.

Kint, N., Janoir, C., Monot, M., Hoys, S., Soutourina, O., Dupuy, B., and Martin-Verstraete, I. (2017). The alternative sigma factor σ^B plays a crucial role in adaptive strategies of Clostridium difficile during gut infection. Environ. Microbiol. *19*, 1933–1958.

Larocque, M., Chénard, T., and Najmanovich, R. (2014). A curated C. difficile strain 630 metabolic network: prediction of essential targets and inhibitors. BMC Syst. Biol. *8*, 117.

Lessa, F.C., Mu, Y., Bamberg, W.M., Beldavs, Z.G., Dumyati, G.K., Dunn, J.R., Farley, M.M., Holzbauer, S.M., Meek, J.I., Phipps, E.C., et al. (2015). Burden of Clostridium difficile infection in the United States. N. Engl. J. Med.372, 825–834.

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

Mani, N., and Dupuy, B. (2001). Regulation of toxin synthesis in Clostridium difficile by an alternative RNA polymerase sigma factor. Proc. Natl. Acad. Sci. USA *98*, 5844–5849.

Martin-Verstraete, I., Peltier, J., and Dupuy, B. (2016). The regulatory networks that control Clostridium difficile toxin synthesis. Toxins (Basel) *8*, 153.

Matamouros, S., England, P., and Dupuy, B. (2007). Clostridium difficile toxin expression is inhibited by the novel regulator TcdC. Mol. Microbiol. *64*, 1274–1288.

McDonald, J.A.K., Mullish, B.H., Pechlivanis, A., Liu, Z., Brignardello, J., Kao, D., Holmes, E., Li, J.V., Clarke, T.B., Thursz, M.R., and Marchesi, J.R. (2018). Inhibiting growth of Clostridioides difficile by restoring valerate, produced by the intestinal microbiota. Gastroenterology *155*, 1495–1507.e15.

Monot, M., Boursaux-Eude, C., Thibonnier, M., Vallenet, D., Moszer, I., Medigue, C., Martin-Verstraete, I., and Dupuy, B. (2011). Reannotation of the genome sequence of Clostridium difficile strain 630. J. Med. Microbiol. *60*, 1193–1199.

Moretto, M., Sonego, P., Dierckxsens, N., Brilli, M., Bianco, L., Ledezma-Tejeida, D., Gama-Castro, S., Galardini, M., Romualdi, C., Laukens, K., et al. (2016). COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. Nucleic Acids Res *44*, D620–D623.

Nawrocki, K.L., Wetzel, D., Jones, J.B., Woods, E.C., and McBride, S.M. (2018). Ethanolamine is a valuable nutrient source that impacts Clostridium difficile pathogenesis. Environ. Microbiol. *20*, 1419–1435.

Neumann-Schaal, M., Metzendorf, N.G., Troitzsch, D., Nuss, A.M., Hofmann, J.D., Beckstette, M., Dersch, P., Otto, A., and Sievers, S. (2018). Tracking gene expression and oxidative damage of O2-stressed Clostridioides difficile by a multi-omics approach. Anaerobe *53*, 94–107.

Ng, K.M., Ferreyra, J.A., Higginbottom, S.K., Lynch, J.B., Kashyap, P.C., Gopinath, S., Naidu, N., Choudhury, B., Weimer, B.C., Monack, D.M., and Sonnenburg, J.L. (2013). Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens. Nature *502*, 96–99.

Nguyen, N.T.T., Contreras-Moreira, B., Castro-Mondragon, J.A., Santana-Garcia, W., Ossio, R., Robles-Espinoza, C.D., Bahin, M., Collombet, S., Vincens, P., Thieffry, D., et al. (2018). RSAT 2018: regulatory sequence analysis tools 20th anniversary. Nucleic Acids Res *46*, W209–W214.

Norsigian, C.J., Danhof, H.A., Brand, C.K., Oezguen, N., Midani, F.S., Palsson, B.O., Savidge, T.C., Britton, R.A., Spinler, J.K., and Monk, J.M. (2020). Systems biology analysis of the Clostridioides difficile core-genome contextualizes microenvironmental evolutionary pressures leading to genotypic and phenotypic divergence. NPJ Syst. Biol. Appl. *6*, 31.

Opdam, S., Richelle, A., Kellman, B., Li, S., Zielinski, D.C., and Lewis, N.E. (2017). A systematic evaluation of methods for tailoring genome-scale metabolic models. Cell Syst *4*, 318–329.e6.

Orth, J.D., Thiele, I., and Palsson, B.Ø. (2010). What is flux balance analysis? Nat. Biotechnol. *28*, 245–248.

Paquette, S.M., Leinonen, K., and Longabaugh, W.J.R. (2016). BioTapestry now provides a web application and improved drawing and layout tools. F1000Res *5*, 39.

Peltier, J., Hamiot, A., Garneau, J.R., Boudry, P., Maikova, A., Hajnsdorf, E., Fortier, L.-C., Dupuy, B., and Soutourina, O. (2020). Type I toxin-antitoxin systems contribute to the maintenance of mobile genetic elements in Clostridioides difficile. Commun. Biol. *3*, 718.

Peterson, E.J.R., Reiss, D.J., Turkarslan, S., Minch, K.J., Rustad, T., Plaisier, C.L., Longabaugh, W.J.R., Sherman, D.R., and Baliga, N.S. (2014). A high-resolution network model for global gene regulation in Mycobacterium tuberculosis. Nucleic Acids Res *42*, 11291–11303.

Pishdadian, K., Fimlaid, K.A., and Shen, A. (2015). SpoIIID-mediated regulation of σK function during Clostridium difficile sporulation. Mol. Microbiol. *95*, 189–208.

Plaisier, C.L., O'Brien, S., Bernard, B., Reynolds, S., Simon, Z., Toledo, C.M., Ding, Y., Reiss, D.J., Paddison, P.J., and Baliga, N.S. (2016). Causal mechanistic regulatory network for glioblastoma deciphered using systems genetics network analysis. Cell Syst *3*, 172–186.

R Core Team (2013). R: a language and environment for statistical computing. R Foundation for Statistical Computing.

Reiss, D.J., Baliga, N.S., and Bonneau, R. (2006). Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC Bioinformatics *7*, 280.

Reiss, D.J., Plaisier, C.L., Wu, W.-J., and Baliga, N.S. (2015). cMonkey2: automated, systematic, integrated detection of co-regulated gene modules for any organism. Nucleic Acids Res *43*, e87.

Riedel, T., Bunk, B., Thürmer, A., Spröer, C., Brzuszkiewicz, E., Abt, B., Gronow, S., Liesegang, H., Daniel, R., and Overmann, J. (2015). Genome resequencing of the virulent and multidrug-resistant reference strain Clostridium difficile 630. Genome Announc *3*, e00276.

Saujet, L., Monot, M., Dupuy, B., Soutourina, O., and Martin-Verstraete, I. (2011). The key sigma factor of transition phase, SigH, controls sporulation, metabolism, and virulence factor expression in Clostridium difficile. J. Bacteriol. *193*, 3186–3196.

Saujet, L., Pereira, F.C., Serrano, M., Soutourina, O., Monot, M., Shelyakin, P.V., Gelfand, M.S., Dupuy, B., Henriques, A.O., and Martin-Verstraete, I. (2013). Genome-wide analysis of cell type-specific gene transcription during spore formation in Clostridium difficile. PLoS Genet *9*, e1003756.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498–2504.

Soutourina, O., Dubois, T., Monot, M., Shelyakin, P.V., Saujet, L., Boudry, P., Gelfand, M.S., Dupuy, B., and Martin-Verstraete, I. (2020). Genome-wide transcription start site mapping and promoter assignments to a sigma factor in the human enteropathogen Clostridioides difficile. Front. Microbiol. *11*, 1939.

Soutourina, O.A., Monot, M., Boudry, P., Saujet, L., Pichon, C., Sismeiro, O., Semenova, E., Severinov, K., Le Bouguenec, C., Coppée, J.-Y., et al. (2013). Genome-wide identification of regulatory RNAs in the human pathogen Clostridium difficile. PLoS Genet *9*, e1003493.

Steglich, M., Hofmann, J.D., Helmecke, J., Sikorski, J., Spröer, C., Riedel, T., Bunk, B., Overmann, J., Neumann-Schaal, M., and Nübel, U. (2018). Convergent loss of ABC transporter genes from Clostridioides difficile genomes is associated with impaired tyrosine uptake and p-cresol production. Front. Microbiol. *9*, 901.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2016). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res *45*, D362–D368.

Tanay, A., Steinfeld, I., Kupiec, M., and Shamir, R. (2005). Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium. Mol. Syst. Biol. *1*, 2005.0002.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., Lomsadze, A., Pruitt, K.D., Borodovsky, M., and Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res *44*, 6614–6624.

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res *45*, D158–D169.

Theriot, C.M., Koenigsknecht, M.J., Carlson, P.E., Jr., Hatton, G.E., Nelson, A.M., Li, B., Huffnagle, G.B., Z Li, J.Z., and Young, V.B. (2014). Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection. Nat. Commun. *5*, 3114.

Underwood, S., Guan, S., Vijayasubhash, V., Baines, S.D., Graham, L., Lewis, R.J., Wilcox, M.H., and Stephenson, K. (2009). Characterization of the sporulation initiation pathway of Clostridium difficile and its role in toxin production. J. Bacteriol. *191*, 7296–7305.

Vallenet, D., Calteau, A., Cruveiller, S., Gachet, M., Lajus, A., Josso, A., Mercier, J., Renaux, A., Rollin, J., Rouy, Z., et al. (2017). MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. Nucleic Acids Res *45*, D517–D528.

Vemuri, R.C., Gundamaraju, R., Shinde, T., and Eri, R. (2017). Therapeutic interventions for gut dysbiosis and related disorders in the elderly: antibiotics, probiotics or faecal microbiota transplantation? Benef. Microbes *8*, 179–192.

Walter, B.M., Rupnik, M., Hodnik, V., Anderluh, G., Dupuy, B., Paulič, N., Žgur-Bertok, D., and Butala, M. (2014). The LexA regulated genes of the Clostridium difficile. BMC Microbiol *14*, 88.

Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res *42*, D581–D591.

Wolfe, C.J., Kohane, I.S., and Butte, A.J. (2005). Systematic survey reveals general applicability of ‘‘guilt-by-association’’ within gene coexpression networks. BMC Bioinformatics *6*, 227.

Woods, E.C., Nawrocki, K.L., Suárez, J.M., and McBride, S.M. (2016). The Clostridium difficile Dlt pathway is controlled by the extracytoplasmic function sigma factor σV in response to lysozyme. Infect. Immun. *84*, 1902–1916.

Please cite this article in press as: Arrieta-Ortiz et al., Predictive regulatory and metabolic network models for systems analysis of *Clostridioides difficile*, Cell Host & Microbe (2021), https://doi.org/10.1016/j.chom.2021.09.008

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
C. difficile transcriptomes included in transcriptional compendium	See Table S1	N/A
Genes associated with *C. difficile* SigB activity	Boekhoud et al., 2020	GEO: GSE152515
C. difficile transcriptome during lactate and glucose supplementation	Hofmann et al., 2021	GEO: GSE149911
C. difficile metabolic model	Kashaf et al., 2017	*i*cdf834
C. difficile Tn-seq gene essentiality	Dembek et al., 2015	N/A
C. difficile genome annotation	Monot et al., 2011	N/A
C. difficile genome annotation	Girinathan et al., 2021	N/A
C. difficile operon predictions	Dehal et al., 2010	http://microbesonline.org/operons/
C. difficile protein-protein interaction network	Szklarczyk et al., 2016	https://string-db.org
C. difficile experimentally supported TF regulons	See Table S2	N/A
Software and algorithms		
R v3.3.3-4.1.0	(R Core Team, 2013)	N/A
Matlab R2019a	https://www.mathworks.com/products/matlab.html	N/A
cMonkey2	Reiss et al., 2015	https://github.com/baliga-lab/cmonkey2
Inferelator	Arrieta-Ortiz et al., 2015	https://github.com/ChristophH/Inferelator
GIMME	Becker and Palsson, 2008	N/A
PRIME	Immanuel et al., 2021	https://github.com/baliga-lab/PRIME
DESeq2	Love et al., 2014	https://github.com/mikelove/DESeq2
MEME suite	Bailey et al., 2015	https://meme-suite.org/meme/
Biotapestry	Paquette et al., 2016	http://www.biotapestry.org
Cytoscape	Shannon et al., 2003	https://cytoscape.org
Drupal	https://www.drupal.org/home	N/A
Adobe Illustrator CS	Adobe Inc	N/A
Inkscape 1.0.2	https://inkscape.org	N/A
Other		
The *C. difficile* Portal	This study	http://networks.systemsbiology.net/cdiff-portal/
R notebook with scripts to perform computational analyses using the reconstructed EGRIN model	This study	https://github.com/marioluisao/Predictive-regulatory-network-models-for-systems-analysis-of-C.-difficile

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Nitin S. Baliga (nitin.baliga@isbscience.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes, publicly available existing transcriptional data. These accession numbers for datasets are listed in the key resources table. The compiled transcriptional compendium is available in the interactive and open access *C. difficile* Web Portal (http://networks.systemsbiology.net/cdiff-portal/).
- All models and supporting resources reported in this study are available in *C. difficile* Web Portal (http://networks.systemsbiology.net/cdiff-portal/).
- Original code is publicly available at the *C. difficile* Web Portal (http://networks.systemsbiology.net/cdiff-portal/). The code can also be found at https://github.com/marioluisao/Predictive-regulatory-network-models-for-systems-analysis-of-C.-difficile.
- Any additional information required to reanalyze data reported in this paper is available from the lead upon request.

METHOD DETAILS

C. difficile genome annotation

An ATCC43255 reference genome was generated and annotated to support *in vivo* transcriptome studies of *C. difficile* per discrepancies noted in the RefSeq genome, particularly among bacteriophage loci and other mobile elements (Girinathan et al., 2021). The updated reference genome was annotated using the NCBI Prokaryotic Genome Automatic Annotation Pipeline (Tatusova et al., 2016), PATRIC (Wattam et al., 2014), and PROKKA (Seemann, 2014) to extract gene features for support of transcriptome pathway enrichment analyses. Bacteriophage loci and genes were identified using PHASTER (Arndt et al., 2016).

C. difficile transcriptional compendium

To generate a transcriptional compendium for *C. difficile*, required for constructing an EGRIN model, a total of 151 publicly available transcriptomes of *C. difficile* 630 (identified by searching for the ‘Clostridioides difficile 630’ term) were downloaded from the NCBI Gene Expression Omnibus (GEO) repository (Barrett et al., 2013) in March 2020. Downloaded transcriptomes were generated by 11 independent studies (Table S1)(Berges et al., 2018; Dineen et al., 2010; Fimlaid et al., 2013; Giordano et al., 2018; Hastie et al., 2018; Ho and Ellermeier, 2015; Hofmann et al., 2018; Janoir et al., 2013; Janvilisri et al., 2010; Ng et al., 2013; Pishdadian et al., 2015). To integrate this data into a single dataset, we computed the $\log_2$ fold-change of each transcriptome with respect to a (study-specific) control condition, as performed in the generation of other transcriptional compendia (Moretto et al., 2016; Peterson et al., 2014; Tanay et al., 2005). This step was not necessary for transcriptional data collected with dual channel arrays that included a normalizing control channel. The resulting transcriptional compendium contained a total of 4,091 gene features and 127 conditions. The 127 conditions in the transcriptional compendium were organized in 11 distinct condition blocks (e.g., sporulation, *fur* deletion), as shown in Table S1. A brief description of each condition included in the final transcriptional compendium is available in the *C. difficile* Portal.

Construction of the EGRIN model

The EGRIN model for *C. difficile* was constructed in two stages. First, we used cMonkey2 (Reiss et al., 2015), a biclustering algorithm, on the compiled compendium of 127 *C. difficile* transcriptomes to simultaneously detect co-regulated gene modules and the conditions where co-regulation occurs. cMonkey2 integrates functional annotation from the STRING database (Szklarczyk et al., 2016), gene promoter sequences from the RSAT database (Nguyen et al., 2018), and operon predictions from MicrobesOnline (Dehal et al., 2010) when detecting gene modules. cMonkey2 was run using default parameters. Briefly, we used 2,000 iterations to optimize the co-regulated gene modules, each one with 3-70 genes. In each iteration, cMonkey2 refined the gene modules by evaluating and modifying (if necessary) condition and gene memberships. cMonkey2 biclustering approach allowed genes and conditions to be assigned to a maximum of two and 204 different modules, respectively. *De novo* GRE search was performed using MEME v. 4.12.0 (Bailey et al., 2015). Second, we used the Inferelator (Arrieta-Ortiz et al., 2015), a network inference algorithm, to identify potential transcriptional regulators for the 406 gene modules generated by cMonkey2. The Inferelator uses a Bayesian Best Subset Regression to estimate the magnitude and sign (activation or repression) of potential interactions between TFs and gene modules based on TF transcriptional profiles and module eigengenes (i.e., first principal component) (Plaisier et al., 2016). We bootstrapped the expression data (100 times) to avoid regression overfitting (Arrieta-Ortiz et al., 2015). The Inferelator generates two scores for each TF-module interaction, the corresponding regression coefficient (i.e., beta) and a confidence score. The second score indicates the likelihood of the interaction. The final set of TF-module interactions was defined as the 805 interactions with absolute beta values ≥ 0.1. Inkscape and Adobe Illustrator were used to generate composite figures.

Literature-derived TF regulons

We mined available literature to compile a list of experimentally supported targets for the 13 partially characterized *C. difficile* transcriptional regulators (involved in sporulation, motility, carbon metabolism, among other processes) shown on Table S2 (Antunes et al., 2011, 2012; Berges et al., 2018; Bouillaut et al., 2019; Dineen et al., 2010; Dubois et al., 2016; Fimlaid et al., 2013; Kint et al., 2017; El Meouche et al., 2013; Saujet et al., 2013, 2011; Soutourina et al., 2020). The manually compiled regulons represented a total of 1,349 regulatory interactions and involved 1,044 genes. Target genes included in the compiled TF regulons were supported by transcriptional data, protein-DNA binding data and *in silico* analysis of promoter regions (e.g., presence of known regulators DNA binding motif).

Please cite this article in press as: Arrieta-Ortiz et al., Predictive regulatory and metabolic network models for systems analysis of *Clostridioides difficile*, Cell Host & Microbe (2021), https://doi.org/10.1016/j.chom.2021.09.008

DNA motif comparison

The MEME suite (Bailey et al., 2015) was used (with default parameters) to reconstruct the known DNA binding motifs of CodY and SigL based on the promoter sequence of their reported target genes (Dineen et al., 2010; Soutourina et al., 2020). Reconstructed TF binding motifs were compared to the GREs associated with modules of interest using Tomtom with default parameters but without scoring of reverse complement sequences.

Function assignment to uncharacterized genes

To predict the potential role of uncharacterized genes, gene functions were predicted based on the functional enrichment of EGRIN modules (evaluated as explained in the ‘Module enrichment evaluation’ section). Function assignments were restricted to uncharacterized genes located in functionally enriched modules in which 45% or more of the annotated members (i.e., genes with putative function) were assigned to the over-represented function term. This second filter was implemented to focus on EGRIN modules that were involved in a specific function. Thus, increasing the likelihood that the uncharacterized genes were also involved in the same function.

Analysis of *in vivo* data

In vivo transcriptomic data from gnotobiotic mice mono-colonized with *C. difficile* ATCC43255 or co-colonized with *P. bifermentans* or *C. sardiniense* (Girinathan et al., 2021) were analyzed as previously described using the updated reference genome of ATCC43255 to extract gene features for subsequent analysis with DESeq2 (Love et al., 2014). For each relevant comparison (e.g., *C. difficile* mono-colonized mice at 24-h vs 20-h of infection) we defined the set of differentially expressed genes (DEGs) as the genes with DESeq2 adjusted p-value < 0.05 and absolute $\log_2$ fold-change > 1. Up- and downregulated genes with orthologs in the CD630 strain were independently mapped into the EGRIN model (as explained below in the ‘Module enrichment evaluation’ section) to identify differentially expressed modules. To gain insights into the adaptation of *C. difficile* to the evaluated *in vivo* conditions, we focused on differentially expressed EGRIN modules that were also enriched with functional pathways or manually-curated TF regulons. We restricted downstream analyses to the enriched modules with absolute median DESeq2 $\log_2$ fold-change ≥ 0.5.

Analysis of transcriptomes not used for training

To illustrate the capability of the EGRIN model to offer functional and mechanistic insights into transcriptional regulation from *C. difficile* transcriptomes that were not included in the compendium used for generating the EGRIN model, we mapped the sets of DEGs upon induced overexpression of SigB (Boekhoud et al., 2020) and low glucose supplementation (Hofmann et al., 2021) into the EGRIN model (Figure S7). EGRIN modules enriched with DEGs were identified as explained above.

Metabolic model refinement

A published genome-scale metabolic model of *C. difficile* 630 strain, icdf834 (Kashaf et al., 2017), was used in this study and expanded by adding reactions required for *in vivo* survival of the pathogen. The icdf834 model incorporates 1,227 metabolic reactions and 807 metabolites. The metabolic reactions were mapped through gene-protein-reaction (GPR) associations to 832 genes, which represent 80% of 1,030 annotated metabolic genes in the CD630 genome (Figure 5A). The original icdf834 model contains two duplicated genes (CD630_28720 and CD630_23220) due to quotation marks in the GPR definitions. We removed these duplicated genes and confirmed that this change did not affect the GPR as the duplicated genes were in an ‘OR’ relationship in the same reaction. We also curated pathway annotations that were incorrectly designated using default KEGG annotations (Kanehisa et al., 2017). For example, most anaerobes do not utilize the tricarboxylic citric acid (TCA) cycle, although some reactions, in reverse, support aspects of pyruvate, succinate and oxaloacete metabolism. In the icdf834 model, we changed subsystem pathway annotation of two reactions - i) acetyl-CoA:oxaloacetate C-acetyltransferase and ii) succinyl-CoA synthase from TCA cycle to pyruvate metabolism and butanoate fermentation respectively (Data S2). Similarly, we updated reactions originally assigned to gluconeogenesis and the pentose phosphate pathway. We evaluated the homology of metabolic genes between *C. difficile* 630 and ATCC43255 strain of *C. difficile* in order to use the icdf834 model for representing the *in vivo* infection state of ATCC43255 strain. The details of 766 genes that are predicted in this homology analysis is provided in Data S2. We then extended the model by adding four genes (CD630_08700, CD630_08680, CD630_17090 and CD630_10810) and eight exchange reactions that are required for the growth of the pathogen in the *in vivo* micro-environment, based on KEGG annotations. We also expanded the proline reductase (PR) and glycine reductase (GR) systems by adding alanine, branched chain amino acids (valine, leucine and isoleucine) and aromatic amino acids (phenyl alanine, tyrosine and tryptophan) as donors in the Stickland metabolism (Jackson et al., 2006), increasing the total number of reactions from 1,227 to 1,273 (Data S2). We named this expanded version of the model as “icdf836”. Then, the transcriptome of *C. difficile* profiled from *in vivo* infections of specifically-colonized gnotobiotic mice (Girinathan et al., 2021) was mapped onto the icdf836 model using the GIMME algorithm with a default threshold (12) for reaction normalized scores, as reported previously (Becker and Palsson, 2008). We decided to use the default value because previous studies have demonstrated that GIMME performance is robust to threshold selection (Becker and Palsson, 2008; Opdam et al., 2017). This resulted in a model with 712 (642) active reactions for the mono-colonized (*P. bifermentans* co-colonized) model, with no changes in the number of genes. This model represents the *in vivo* state of *C. difficile*. We applied the constraint-based method for simulating the metabolic steady-state of *C. difficile* using flux-balance analysis (FBA) (Becker and Palsson, 2008; Orth et al., 2010). The initial validation steps involved checking the capacity of the icdf834 model to produce biomass in defined media conditions including 1) minimal medium, 2) basal defined medium and 3) complex, nutrient-rich medium (compositions used according to Larocque et al., 2014). Then, we tested the performance of both icdf834 and icdf836 models using gene essentiality predictions by FBA.

Gene essentiality prediction

A gene was considered "essential" if its *in silico* deletion in the metabolic model reduced the biomass by >65%. By this analysis, the model classified each gene as "essential" or "non-essential". We compared the gene essentiality predictions from nutrient-rich media constraints with the available experimental Tn-seq data (Dembek et al., 2015) and deduced the confusion matrix to derive true positive rates (TPR) and false positive rates (FPR). This led to the elucidation of sensitivity and specificity of the model using ROC curve analysis. We then applied the same strategy and predicted the essential genes *in vivo* using FBA with the expanded and GIMME-derived context-specific network, icdf836. All model simulations related to FBA were performed on MATLAB_R2019a platform using the recent version of COBRA (The COnstraint-Based Reconstruction and Analysis) toolbox (Heirendt et al., 2019). *In silico* gene essentiality predictions were performed using the COBRA toolbox 'single-gene-deletion' function in MATLAB.

PRIME model development

Three PRIME models were constructed for *C. difficile* during *in vivo* (mono- and *P. bifermentans* co-colonized mice) and *in vitro* growth. Briefly, we first inferred a TF-gene transcriptional network using the Inferelator with incorporation of TF activity, estimated using the signed (positive or negative) TF-gene interaction network compiled for 13 TFs (Table S2), as we have previously done for other species (Arrieta-Ortiz et al., 2015, 2020). To infer a global network, 288 putative transcriptional regulators without known targets were also included as potential regulators. The inferred network included 5,801 TF-gene interactions with absolute regression coefficients ≥ 0.1. The transcriptional network was then integrated with the icdf836 metabolic model as explained in (Immanuel et al., 2021). Importantly, PRIME models were made context-specific by excluding metabolic reactions associated with lowly expressed genes, as explained above.

Network visualization

Illustration of transcriptional and metabolic networks were generated using BioTapestry (Paquette et al., 2016) and Cytoscape (Shannon et al., 2003).

QUANTIFICATION AND STATISTICAL ANALYSIS

Module enrichment evaluation

We used a hypergeometric test to identify modules of co-regulated genes in the EGRIN model that were statistically enriched with manually compiled TF regulons (Table S2) or functional pathways derived from curated annotation of *C. difficile* genome (Girinathan et al., 2021). Only gene modules with adjusted hypergeometric test p-value ≤ 0.05 and containing five or more genes from the relevant TF regulon or functional pathway were considered enriched. The same approach was used to identify EGRIN modules enriched with any gene set of interest.

Transcriptional profiles of EGRIN modules

To identify EGRIN modules that were responsive (i.e., up- or down-regulated) to a particular perturbation (i.e., condition block) included in the compiled transcriptional compendium (Table S1), for each module we computed condition-wise median of the $\log_2$ ratios of all genes in the module. Then, the set of conditions assigned (during biclustering) to a module was organized in ascending order based on the computed median values, and divided in quintiles. A hypergeometric test was used to identify condition blocks over-represented in the first and fifth quintiles, which correspond to the ones with the lowest and highest $\log_2$ ratios, respectively. Hypergeometric test p-values ≤ 0.001 were considered significant. To filter out modules with inconsistent fold-changes, only modules enriched with condition blocks in their first (or fifth) quintile and average $\log_2$ ratios < -1 (or average $\log_2$ ratios > 1) were considered differentially active and included in Data S1. A total of 261 instances of differential activity, involving 218 modules, were identified.

To evaluate the probability of observing modules with differential activity (as defined previously) by chance, we generated 1,000,000 biclusters by randomly sampling from genes and conditions in the transcriptional compendium. The number of genes in each bicluster was defined by randomly sampling the distribution of gene counts for the EGRIN modules. A similar approach was used to define the number of conditions in each random bicluster. Overall, we observed 66,977 instances of differential activity (involving 64,038 random biclusters). Remarkably, the proportion of EGRIN biclusters with differential activity (53.7%) is significantly higher (hypergeometric test p-value < 1e-145 based on the estimated null distribution) than the proportion of modules with differential activity in the set of random biclusters (6.4%).

Please cite this article in press as: Arrieta-Ortiz et al., Predictive regulatory and metabolic network models for systems analysis of *Clostridioides difficile*, Cell Host & Microbe (2021), https://doi.org/10.1016/j.chom.2021.09.008

ADDITIONAL RESOURCES

The *C. difficile* web portal

The *C. difficile* portal (http://networks.systemsbiology.net/cdiff-portal/) utilizes the powerful build, search, collaboration, and visualization features of the Drupal content management system. Leveraging Drupal's modularity and extensibility, we developed this content management system into a data management, analysis, and visualization framework to support *C. difficile* research.

Due to the complexity of the information provided by the genome and models, it is critical to provide a user-friendly and flexible search and filtering capabilities. By taking advantage of Drupal's built-in search interface and implementing Apache Solr search, the portal database includes the capability to search by facets, which together with sorting enables users to start with general searches and then quickly pinpoint specific information.

In order to provide a comprehensive functional genomics resource for the *C. difficile* community, genome annotations from several different sources were merged and imported into the *C. difficile* Portal. Curated genome annotations for *C. difficile* strain 630 published by Monot et al. (Monot et al., 2011), were downloaded from MicroScope platform (Vallenet et al., 2017). Additional functional annotations were downloaded from PATRIC (Wattam et al., 2014) and Uniprot (Consortium, 2017) and merged with curated genome annotations. Overall, 4,018 genes were included in the *C. difficile* Portal. The *C. difficile* genome included 1,030 metabolic genes, 309 TFs, 270 small non-coding RNAs (sRNAs) (Soutourina et al., 2013), 87 tRNAs, 32 rRNAs and 17 miscellaneous RNAs (miscRNAs). The genome included 1,330 genes with unknown function. Furthermore, gene essentiality data from Dembek et al. (Dembek et al., 2015) was integrated with gene annotations.