# 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], Jay N. Worley[2], Nicholas DiBenedetto[2], Olga Soutourina[3], Johann Peltier[3], Bruno Dupuy[4], Lynn Bry[2] and Nitin S. Baliga[1,+]

1. Institute for Systems Biology, Seattle, WA, USA
2. Massachusetts Host-Microbiome Center, Dept. Pathology, Brigham & Women's Hospital, Harvard Medical School, Boston, MA , USA
3. Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France
4. Laboratory of the Pathogenesis of Bacterial Anaerobes, Institut Pasteur, Université de Paris, UMR-CNRS2001, France
+corresponding author: nitin.baliga@isbscience.org

# ABSTRACT

Though *Clostridioides difficile* is among the most studied anaerobes, we know little about the systems level interplay of metabolism and regulation that underlies its ability to negotiate complex immune and commensal interactions while colonizing the human gut. We have compiled publicly available resources, generated through decades of work by the research community, into two models and a portal to support comprehensive systems analysis of *C. difficile*. First, by compiling a compendium of 148 transcriptomes from 11 studies we have generated an **E**nvironment and **G**ene **R**egulatory **I**nfluence **N**etwork (EGRIN) model that organizes 90% of all genes in the *C. difficile* genome into 297 high quality modules based on evidence for their conditional co-regulation by at least 120 transcription factors. EGRIN predictions, validated with independently-generated datasets, have recapitulated previously characterized *C. difficile* regulons of key transcriptional regulators, refined and extended membership of genes within regulons, and implicated new genes for sporulation, carbohydrate transport and metabolism. Findings further predict pathogen behaviors in *in vivo* colonization, and interactions with beneficial and detrimental commensals. Second, by advancing a constraints-based metabolic model, we have discovered that 15 amino acids, diverse carbohydrates, and 24 genes across glyoxylate, Wood-Ljungdahl, nucleotide, amino acid, and carbohydrate metabolism are essential to support growth of *C. difficile* within an intestinal environment. Models and supporting resources are accessible through an interactive web portal (http://networks.systemsbiology.net/cdiff-portal/) to support collaborative systems analyses of *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 (1). 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 creates altered nutrient states in intestinal environments due to lack of competition for nutrients from host, dietary or microbial origin. The pathogen modifies its endogenous metabolism to respond to these altered states, which stimulates subsequent cellular programs that can promote enhanced colonization and growth. Stress and starvation responses within *C. difficile* populations trigger responses that lead to sporulation, biofilm formation and release of mucosal damaging toxins (2–4).

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. The PaLoc also contains *tcdR*, a sigma factor specific for the toxin gene promoters, and *tcdC*, a TcdR anti-sigma factor (5, 6). Epidemic ribotype 027 strains carry a second toxin locus, *cdt,* which includes the binary toxin genes *ctdA* and *ctdB*, and *cdtR* regulator, which has also been hypothesized to modulate PaLoc expression (7, 8). *C. difficile* elaborates toxin under starvation conditions to extract nutrients from the host and promote spore shedding (9–11). Regulation of PaLoc expression occurs via a complex network of TFs and small molecule inputs, of which direct primary regulators have been well described, but more complex and combinatorial effects remain unclear (11). Toxin production further triggers rapid and profound host immune responses, including release of reactive oxygen species (ROS) which substantially alters the redox state of the gut environment, and other innate immune responses that can induce *C. difficile* stress responses to cell wall, oxidative, and other damaging stimuli (12–15). As per all microbes, *C. difficile* adapts to complex, dynamic environments through

64 changes in metabolism coordinated by a gene regulatory network (16, 17). However, the
65 mechanisms by which the gene regulatory network and metabolic pathways integrate to modulate
66 *C. difficile* pathogenesis remain ill-defined (18, 19).

67 The *C. difficile* 630 (CD630) genome encodes 4,018 genes, with ~309 putative transcription
68 factors (TFs; including sigma factors), 1,030 metabolic genes, and 1,330 genes (>30%) with
69 unknown function (20, 21). The ATCC43255 strain of *C. difficile,* which is used to generate
70 symptomatic infections in mice, encodes 4,117 genes and ~327 TFs, of which ~97% are
71 significantly orthologous to genes encoded in the CD630 strain (22). To address questions
72 regarding the broader systems-level interplay among genes in colonization and infection, we used
73 computational modeling and network inference algorithms to construct an Environment and Gene
74 Regulatory Influence Network (EGRIN) model for *C. difficile.* This model leverages a compendium
75 of 148 published transcriptomes that surveyed responses of CD630 in diverse contexts. The
76 EGRIN model consists of modules of putatively co-regulated genes identified based on their co-
77 expression over subsets of conditions, enrichment of functional associations, chromosomal
78 proximity, co-occurrence across phylogenetically related organisms, and presence of conserved
79 DNA motif(s) within their promoter regions indicating regulation by the same TFs. Further, using
80 regression analysis, EGRIN also captures the combinatorial regulation of genes within each
81 module as a function of the weighted influences of TFs. The model supports a systems-level
82 understanding of the infective capacity of this obligate anaerobe under different *in vitro* and *in vivo*
83 conditions.

84 In addition to EGRIN, we have advanced a metabolic network model of *C. difficile* to understand
85 how conditional regulation manifests physiologically, by adding reactions and associated genes
86 supporting the exchange of nutrients required for growth in intestinal environments. We apply the
87 EGRIN and metabolic models to predict conditional contributions of metabolic genes to the
88 pathogen's fitness under different environmental conditions. Analyses support rational prediction

of context-specific vulnerabilities of the pathogen and uncover TFs driving essential adaptive responses under *in vitro* versus *in vivo* conditions. This analytic framework provides a new 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* (23, 24). 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 148 publicly available transcriptomic datasets from 11 independent studies using CD630 (**Table 1**). This compendium captures diverse transcriptional responses of *C. difficile* to commensals, *in vitro* or *in vivo* responses to different nutrient conditions, and consequences of targeted TF gene deletions. The transcriptome compendium together with functional associations from STRING (25), and all promoter sequences, was analyzed with a suite of network inference tools to infer an EGRIN model for *C. difficile* (**Fig. 1A**). The cMonkey2 biclustering algorithm (26) iteratively grouped functionally-associated genes into **modules** based on their co-expression across subsets of environments, and presence of similar cis-acting gene regulatory elements (**GREs**), providing mechanistic evidence for co-regulation. Subsequently, we used the Inferelator (27) to discover potential TFs of each module through a regression-based approach. The resulting EGRIN model organized 3,895 of 4,018 CD630 genes into 406 gene modules, and inferred module regulation by 148 of 309 genomically identified TFs that putatively act through GREs discovered within gene and operon promoters. Among the Inferelator implicated regulatory networks, 221 modules were controlled by more than one TF, and 75 were regulated by more than two TFs (**Fig. S1**). The TF module assignments support subsequent hypothesis-driven experimental analyses, including the design of ChIP-seq and TF-deletion experiments to validate the regulatory network architecture under physiologically relevant environmental contexts.

The quality of modules within the EGRIN model was evaluated using residual scores, which reflect the coherence of gene co-expression patterns. The lower the residual score, the higher the quality of the module. We determined using an empirical approach that 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 (**Fig. 1B**). Incidentally, this empirically determined threshold was similar to the threshold used to identify high quality EGRIN gene modules for *Mycobacterium tuberculosis* (28). Altogether, the 297 high quality modules captured transcriptional regulation of 3,617 genes (90%) in CD630, with average membership of 20 genes per module (**Fig. 1C-D**). These metrics were consistent with models developed for other organisms (28, 29), a remarkable finding given that the transcriptional dataset used to construct the *C. difficile* model was less than 10% the size of ones 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. To do so, we performed gene enrichment analysis within modules using an updated annotation of *C. difficile* genome (22). This analysis identified 93 of 297 modules (31%) with significant enrichment of genes with related functions in 40 pathways (hypergeometric test adjusted p-value ≤ 0.05). Among these pathways, 13 were over-represented in three or more modules (**Fig. 2A**), demonstrating the capacity of the model to discover conditional partition of cellular processes. We also investigated whether the EGRIN model had identified known regulatory interactions between TFs and their target genes. We compiled from literature the regulons (i.e. target genes) of 13 previously characterized TFs in *C. difficile*, representing a network of 1,353 TF-gene interactions (**Table S1**). Notably, a total of 65 modules (22% of all high-quality modules) were significantly enriched with nine of these TF regulons (**Fig. 2B**). The EGRIN model recapitulated 541 of the 1,212 (45%) previously characterized interactions. This value is consistent with the recall rate of the EGRIN model for *M.*

154 *tuberculosis* (41%-49%) (28). The poor recall of the remaining four regulons (141 regulatory
155 interactions) could be due to underrepresentation of gene expression data from relevant
156 conditions in which these regulons are conditionally active. This analysis also uncovered
157 combinatorial regulation of genes across 19 modules (i.e. enriched with more than one TF
158 regulon). Consistent with the known hierarchical scheme for regulation of sporulation (30),
159 expression of 161 genes across at least eight modules were putatively influenced by Spo0A in
160 combination with one or more alternative sigma factors implicated in sporulation (e.g., SigE).
161 EGRIN also predicted CcpA contributions in seven additional modules in combination with CodY,
162 PrdR and SigL, illustrating the complexity of modular transcriptional regulation in *C. difficile*.
163
164 The biclustering of genes by cMonkey2 is constrained by the *de novo* discovery of a conserved
165 GRE(s) within their promoters in order to cluster genes that are co-regulated, and not just co-
166 expressed. The GREs represent putative binding sites for TFs that are often independently
167 implicated by the Inferelator and protein-DNA interaction maps as regulators of genes within the
168 same module(28, 29, 31). Notably, we determined that the GREs within promoters of genes in
169 modules #182 and #309 were similar to known binding sites for CodY and SigL, which are
170 predicted regulators of those modules, and are among the very few TFs for which binding sites
171 have been characterized (**Fig. 2C-D**).
172
173 The EGRIN modules also detected co-regulation of genes within and across functionally related
174 operons. For example, module #152, which is enriched with the SigD regulon, contains 16 genes
175 that were part of four operons including the flagellar operon *flgG1G-fliMN*-CD630_02720-*htpG,* in
176 addition to *pyrBKDE*, CD630_30270-CD630_30280-*malY*-CD630_30300, and CD630_32430-
177 *prdA*. *De novo* search for TF motifs traditionally use dozens of sequences to identify putative
178 GREs. With the amount of transcriptomic information available, further robust prediction of
179 putative GREs was limited by available numbers of putative binding sites after genes were

organized into predicted operon structures. In addition, multiple modules with statistically significant GREs could not be matched to characterized TFs due to the limited number of TFs with known motifs in *C. difficile*. These limitations can be overcome with additional transcriptomic datasets, as leveraged for EGRIN model development for other species.

### ***C. difficile*** **EGRIN model uncovers regulatory networks for the Pathogenicity Locus**

We evaluated capacity for the EGRIN model to recall known mechanisms of PaLoc regulation, and to provide new information regarding complex regulatory and small molecule effects. The EGRIN model captured certain previously described effects of CodY on toxin gene expression (**Fig. 2E**), as shown in module #182, which is enriched with members of the CodY regulon including *tcdA*. In agreement with EGRIN-predicted CodY regulation of PaLoc genes in module #182, genes encoding the toxin *tcdA* and its regulator *tcdR* were significantly overexpressed upon deletion of *codY* (**Fig. 2G**). Interestingly, *tcdB* which was co-regulated with sporulation genes in module #397 was also significantly upregulated in the *codY* deletion strain, suggesting that this effect might be an indirect consequence of disrupted CodY regulation of *tcdR* (**Fig. 2G**). It is assumed that CodY acts on PaLoc gene expression primarily through its repression of *tcdR.* Putative lower affinity binding sites have been suggested in the toxin gene promoter regions (32). The presence of the CodY motif (**Fig. 2C**) in most members of module #182, including *tcdA* (purple font in **Fig. 2E**) suggests direct influence of CodY on *tcdA* gene expression. The EGRIN model also identified previously reported connections between sporulation and toxin production (33). *tcdB* was assigned to module #397, which was significantly enriched with genes controlled by Spo0A, the master regulator of sporulation (**Fig. 2F**). Additional members of the PaLoc were assigned to other modules, supporting the presence of multiple condition-dependent promoters within the PaLoc (**Table S2**).

### **Assignment of putative functions to genes in EGRIN modules**

Approximately 33% of gene features in the CD630 genome have unknown functions. Thus, the *C. difficile* EGRIN model emerges as a resource to assign putative functions to uncharacterized genes based on functional associations among co-regulated genes (i.e. guilt-by-association)(34). We predicted putative functions for 48 uncharacterized genes by mining underlying functional enrichment of modules under different experimental conditions (see methods). These 48 previously uncharacterized genes were associated with 13 functional categories, including "Sporulation" and "Other sugar-family transporters" (**Fig. 3A**).

Ten genes were putatively assigned sporulation-related functions based on their co-regulation in the sporulation associated modules #206 and #251 (**Fig. 3B)**. Module #251 includes the sporulation-associated alternative sigma factors SigG and SigE (located in the same operon). Module #206 includes seven stage III sporulation genes (*spoIIIAA*, *spoIIIAB*, *spoIIIAC*, *spoIIIAD*, *spoIIIAF*, *spoIIIAG* and *spoIIIAH*), and two stage IV sporulation genes (*spoIV*, *spoIVA*). Reduced expression of the 10 putative sporulation genes upon deletion of sporulation-associated sigma factors suggested putative roles within the mother cell or the forespore. Seven genes are likely associated with mother cell-specific roles based on their decreased expression in *sigE* (six genes) and *sigK* (one gene) deletion strains (**Table S1**). Two additional genes were down-regulated in a *sigG* deletion strain, suggesting putative functions in the forespore. Notably, Tn-seq studies for gene essentiality in *C. difficile* identified seven of these 10 genes as required for sporulation (35).

Module #43 contains six genes, organized in a single operon (CD630_15840-15890), associated with the category "Other Sugar-family transporters". Studies by Antunes et al. (3) identified members of the CD630_15840-15890 operon to be regulated by glucose and indirectly by CcpA. Thus, the 12 uncharacterized genes included in module #43 may also be associated with the same functional category (**Fig. 3C**). Two of these uncharacterized genes (CD630_13011 and CD630_29661) were also identified as CcpA targets in the presence of glucose (3). These

EGRIN-predicted functional assignments are consistent with the known role of CcpA in regulating sugar transport and metabolism (3).

Module #48 contains two adjacent operons (*4hbd-cat2*-CD630_23400-*abf2* and *sucD-cat1*) associated with aminobutanoate degradation. Both operons are regulated by CodY and PrdR. Hence, we predicted that the four uncharacterized genes in this module may be also involved in amino acid metabolism (**Fig. 3C**). In support of this hypothesis, CD630_08760 and CD630_08780 are both differentially expressed upon *codY* deletion. Recent studies also suggest that CD630_08760 may function as a tyrosine transporter per its homology to the CodY-regulated neighbor gene, CD630_08730 (36). Furthermore, Steglich et al. (37) observed decreases in tyrosine uptake and Stickland fermentation in clinical isolates lacking CD630_08760 and CD630_08780.

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

We investigated the differential expression of EGRIN modules across multiple published *in vivo* experiments to discover underlying regulatory mechanisms that drive *C. difficile's* colonization and adaption to *in vivo* environments. This analysis discovered the *in vivo* activation of module #158, particularly during acute infection, and the down regulation of module #48; notably the latter was upregulated during early infection (**Fig. 4A-B**). Module #48 is enriched with members of the CodY and PrdR regulons, as described above. Module #158 is enriched for putative PrdR and EutV co-regulated ethanolamine utilization genes, including *eut* operons for a 2-component histidine kinase sensing system and carboxysome structural proteins that house the ethanolamine fermentative enzymes (38). Ethanolamine is prevalent within gut secretions and is also released from damaged host tissues, providing a readily available carbon and nitrogen source for *C. difficile*. The predicted co-regulation of this gene module by PrdR suggests additional *in vivo* functions of this regulator to optimize *C. difficile's* metabolism in gut environments.

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 that were mono-colonized with the mouse-infective strain *C. difficile* ATCC43255 or co-colonized with *C. difficile* and the protective gut commensal species *Paraclostridium bifermentans* (PBI), or infection-worsening species *Clostridium sardiniense* (CSAR). These datasets were not used in model construction. By mapping sets of differentially expressed genes into the EGRIN model we uncovered modules across 20 cellular processes and their associated TFs that were differentially regulated in the presence of PBI or CSAR (**Fig. 4C-F**).

Two sporulation-enriched modules (modules #206 and #261) were up-regulated by 24h of infection in monocolonized mice (**Fig. 4C**). The same two modules were up-regulated by 24h of infection in CSAR co-colonized mice, in addition to four other modules also enriched with sporulation genes (modules #82, #223, #242, #251 in **Fig. 4D**). These six modules were enriched with the Spo0A regulon. On the other hand, no sporulation-enriched modules were detected by 24h of infection in PBI co-colonized mice (**Fig. 4E**). Comparison of CSAR co-colonized mice and PBI co-colonized mice discovered four sporulation-enriched modules (including modules #206 and #261) with increased expression in the virulent context (i.e. presence of CSAR) (**Fig. 4F**). These findings were confirmed with the high levels of spore release in expanded populations of vegetative *C. difficile* when co-colonized with CSAR (22). Overall, 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 (**Fig. 2F**).

Module #319 contains multiple genes associated with electron transport via Rnf ferredoxin systems, and steps in glycolytic, butanoate and succinate metabolic pathways. This module was upregulated at later stages of infection in monocolonized mice at 24h (**Fig. 4C**), and was also up-

284 regulated in CSAR co-colonized mice when compared to PBI co-colonized mice (**Fig. 4F**). Module
285 #319 was consistently down-regulated in mice co-colonized with the protective commensal PBI
286 (**Fig. 4E**). These findings show associated activation of multiple co-regulated energy generating
287 pathways in hypervirulent states of *C. difficile*. Because the EGRIN model identified the
288 NAD+/NADH sensing regulator Rex as a potential activator of module #319, the observed down-
289 regulation of module #319 in PBI co-colonized mice indicates decreased Rex activity. This may
290 explain why a *rex* deletion strain supported increased survival in hamsters (24).
291
292 Five modules enriched with the SigD-regulated genes encoding subunits of flagella (modules
293 #184, #187, #295, #296 and #358) were downregulated in monocolonized mice at 24h (**Fig. 4C**).
294 Similarly, two modules enriched with SigD-regulated motility genes (modules #152 and #295)
295 were downregulated in CSAR co-colonized mice (**Fig. 4D**). From these modules, only module
296 #152 was downregulated in PBI co-colonized mice (**Fig. 4E**), indicating that motility may be
297 repressed to redirect resources toward pathogenesis. This finding is supported by increased
298 virulence of *C. difficile* strains lacking a functional flagella (39). Surprisingly, module #273
299 enriched with the SigD regulon but not with flagellar genes was downregulated in PBI co-
300 colonized mice (**Fig. 4E**) but upregulated in CSAR co-colonized mice (**Fig. 4F**). One of the genes
301 in this module *luxS* encodes a protein involved in the synthesis of the quorum sensing signal, and
302 its over-expression increases toxin expression (40). While it is unclear whether SigD plays a role
303 in the expression of this module, these observations suggest that downregulation of module #
304 273 and *luxS* (through a still uncharacterized mechanism) may contribute to the PBI-mediated
305 reduction of *C. difficile* virulence. In summary, the described EGRIN modules illustrate the
306 potential of the model to uncover additional co-regulated genes and cellular functions that enable
307 states of enhanced virulence within *C. difficile*, and support multiple additional hypotheses for
308 experimental validation.
309

**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 leveraged reconstructed metabolic models that mapped functionally annotated genes to curated biochemical reactions. We extended a previously developed icdf834 metabolic model for *C. difficile* strain 630 (41, 42). The icdf834 model incorporates 1227 metabolic reactions and 807 metabolites. The metabolic reactions were mapped through gene-protein-reaction (GPR) associations to 834 genes, which represent 80% of 1,030 identified metabolic genes in the CD630 genome (**Fig. 5A**). We increased the number of genes in the icdf834 model from 834 to 838 (**Fig. 5B**), and added six new exchange reactions to account for *C. difficile's* capacity to utilize mannitol, fructose, sorbitol, raffinose, succinate and butanoate (43, 44) (**Fig. 5B**). We also added four genes (CD630_08700, CD630_08680, CD630_17090 and CD630_10810) that encode three reactions for molybdenum utilization and cofactor synthesis. Lastly, we updated pathway annotations to reflect those found in obligate anaerobes. For example, the tricarboxylic citric acid cycle (TCA) is not found in most anaerobes, though 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 (**Supplementary File S1**). Similar updates were performed for reactions originally assigned to gluconeogenesis and the pentose phosphate pathway. We refer to this updated model as icdf838. Lastly, the model derived from CD630 was compared with the gene feature content from *C. difficile* ATCC43255, used commonly in mouse infection models. The two strains shared 92% of metabolic genes and predicted pathways (768 out of 838 genes in the icdf838 model have homology with the ATCC43255 strain; **Supplementary File S1**).

We validated the completeness and accuracy of this model by confirming its ability to predict biomass production in three different *in vitro* media compositions: 1) minimal medium, 2) basal defined medium and 3) complex, nutrient-rich medium (see Larocque et al. 2014 (42) for media compositions). The model accurately predicted *C. difficile's* requirements for six amino acids: cysteine, leucine, Isoleucine, proline, tryptophan and valine (45). We also tested the performance of this "*in silico broth*" model for accuracy in predicting gene essentiality by comparing our model predictions to results from Tn-seq fitness screen performed *in vitro* under nutrient-rich conditions (35). With a threshold cutoff of 95% predicted growth inhibition, receiver-operator curve (ROC) analyses demonstrated high sensitivity and specificity of the model predictions (**Fig. 5C**; area under curve = 0.7626; p-value=0.015), indicating capacity for the model to distinguish essential versus non-essential gene calls with a true positive rate (sensitivity) of 0.9791 and a false positive rate (specificity) of 0.5431(**Fig. 5C**).

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 (22) were used as input into the GIMME algorithm (46). Analyses of expressed transcripts *in vivo* identified 665 active reactions (**Fig. 5B**) within the icdf838 model during colonization, growth, and over the course of symptomatic infection. Leveraging information from the *in vitro* studies, the model made two notable predictions *in vivo* regarding the pathogen's metabolism. First, the icdf838 model predicted 15 amino acids to be required for *C. difficile* growth in contrast to the 6 required *in vitro* (**Supplementary File S1**). These amino acids included the dominant Stickland-fermented amino acids that were also required *in vitro,* including proline and branched chain amino acids, and additional amino acids including arginine, glutamate, lysine and methionine, which also have multiple cellular functions in cell wall synthesis, nitrogen cycling, and responses to oxidative stress. Secondly, the model predicted *C. difficile's* switch from preferential use of glucose as a carbon source *in vitro* in complex media, to simultaneous utilization *in vivo* of diverse carbohydrate

sources including fructose, galactose, maltose, and sugar alcohols such as mannitol and sorbitol, to promote colonization and growth (**Supplementary File S1**). Seven of these carbohydrate sources were described in other *in vivo* mouse infection studies illustrating support for these findings across *C. difficile* strains, and in germfree and conventional mouse models (**Supplementary File S1**) (43, 44, 47).

We next used the metabolic model to identify essential metabolic genes and networks that promote *C. difficile's* growth *in vivo*. Gene deletions predicted to reduce the pathogen's *in vivo* growth by ≥95% identified 24 genes, involved in 1 carbon-cycling reactions in glyoxylate and Wood-Ljungdahl metabolism, nucleotide biosynthesis, nucleotide interconversion and salvage pathways, amino acid biosynthetic and metabolic reactions, and aspects of central carbohydrate metabolism (**Fig. 5D-E**). These metabolic pathways represent new potential targets that drive aspects of *C. difficile's* colonization and subsequent growth which are required to develop symptomatic infections. Model predictions also illustrated *C. difficile's* predicted shift from carbohydrate utilization towards amino acid utilizing pathways *in vivo*, as shown by the enhanced set of 15 amino acids, including the preferred Stickland donor and acceptor amino acids (leucine and proline) known to support metabolism and growth (43, 44, 47). Notably, many of these amino acids show high abundance within the gut lumen in gnotobiotic and conventional colonization states that enhance *C. difficile's* capacity to colonize and expand (22).

**The Cdiff Web Portal, a resource for the *C. difficile* community**

We have released a new *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, gene regulation, and metabolism on its virulence. Within the portal users can access publicly

386 available datasets (e.g. transcriptional compendia), models, software and supporting resources.
387 The Portal includes information on more than 4,000 *C. difficile* genes, 1,227 metabolic reactions,
388 and 406 co-regulated gene modules. Genes can be explored in the context of genome
389 annotations, expression profiles, regulatory and metabolic membership, and other functional
390 genomic information across different databases including COG, Uniprot, and PATRIC (48–50).
391 The portal provides access to detailed information on (1) Genes, (2) predicted Gene Modules,
392 and (3) Metabolic Reactions (**Fig. S2**).
393
394 Each gene module page includes summary statistics for the module, expression profiles of the
395 module genes across conditions incorporated in developing the model, regulatory motifs,
396 regulatory influences from transcription factors, functional enrichment information, and
397 information about regulon member genes (**Fig. 6**). The module pages are structured to facilitate
398 the assessment of the quality and statistical significance of the modules and highlight functional
399 connections. The portal includes a table of metabolic reactions with details of each reaction,
400 associated genes, metabolites, and sub-systems. Metabolites and sub-systems are defined as
401 taxonomic vocabularies that collect and group associated reactions to identify related metabolic
402 processes. In addition, the portal provides access to algorithms, software, and data, and will
403 include information about animal models, strains, and other *C. difficile* relevant community
404 resources. As additional datasets are communicated, model predictions and tools will be
405 successively enhanced to support systems-level analyses and assist in hypothesis generation in
406 *C. difficile* biology and to enable tangible clinical interventions.
407

# DISCUSSION

The obligate anaerobe *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 *in vivo*. 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, the first 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. Recent improvements in the genetic manipulation of *C. difficile*, including the mouse infective strain ATCC43255, open new capacity to probe the GREs modulating critical aspects of its metabolism, growth and virulence, from a systems-level perspective.

The *C. difficile* EGRIN model enables a number of predictions relevant to *in vivo* disease. For example, the PrdR regulator of the pathogen's Stickland proline reductase (*prd*) and other genes, has long been hypothesized to have a role in PaLoc gene expression through as-yet unknown mechanisms. EGRIN predictions included gene regulatory module #182 which identified combined PrdR and CodY effects on *tcdA* gene expression, providing a regulatory integration point and broader set of co-regulated genes to support further experimental analyses of co-regulation between these two transcription factors, including effects on PaLoc expression. 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 new information to support further experimental

433 investigation of combinatorial effects of these and other regulators identified in PaLoc gene-
434 associated biclusters. The EGRIN model also predicted PrdR as a critical regulator *in vivo* through
435 its systems-level effects on the pathogen's colonization, metabolism and growth, involving
436 multiple direct and indirect effects upon other modules and aspects of the pathogen's metabolism
437 and gene regulation.
438
439 The present model did not identify all experimentally known regulators of PaLoc expression,
440 including SigD regulation of TcdR, and effects of other more recently identified PaLoc regulators
441 such as RstA and LexA, for which limited datasets exist from targeted deletion mutants or under
442 multiple nutrient and other environmental perturbations. Nonetheless, as shown with our *in vivo*
443 analyses, application of the EGRIN and metabolic models to new datasets offers key insights into
444 causal mechanistic drivers of adaptive strategies of the pathogen. Given that less than 10% of
445 transcriptomic information and less than 1% of ChIP-seq regulator datasets were available for *C.*
446 *difficile* 630, as compared to EGRIN models developed for other species, the model provides a
447 formative tool to design future transcriptomic and ChIP-seq studies to improve predictions for
448 these regulons.
449
450 Leveraging additional Tn-seq and *in vivo* transcriptomic datasets, the expanded icdf838 model
451 identified a broader set of amino acids, in addition to genes and anaerobe-specific pathways,
452 needed to support colonization and growth expansion *in vivo*. Notably, predictions of *in vivo* gene
453 essentially identified multiple genes in glyoxylate metabolism, a pathway essential in many
454 acetogenic anaerobes (51, 52) that leverage this system with folate 1-carbon cycling pathways
455 including those connected with Wood-Ljundahl fixation of carbon dioxide to acetate. Predictions
456 of gene essentiality also identified multiple nucleotide synthesis and salvage pathway genes that
457 were essential *in vivo* but not *in vitro,* including ones associated with xanthine transport and
458 metabolism, an abundant nucleotide in gut secretions that originates from host sources (22).

Lastly, predictions identified genes in amino acid biosynthetic pathways for branched-chain amino acids, aromatic amino acids, and others that were predicted to be required *in vivo.* Each of these provides new targets of vulnerability for which to consider therapeutic interventions leveraging small molecules, bacteriotherapeutic, or other patient interventions.

We illustrate additional predictions from the *C. difficile* EGRIN model to enable gene- through systems-level analyses of the pathogen. Though among the best described obligate anaerobes, the *C. difficile* genome still contains a high number of genes of unknown function. Model predictions provided new information to assign putative functions to 48 gene features, including ones associated with sporulation, carbohydrate transport, and other aspects of cellular metabolism. The *C. difficile* Web Portal, makes these tools and resources available to the broader *C. difficile,* microbiology, and systems biology communities, providing a platform for collaboration and to support systems-level investigations of the pathogen and its interactions with the host and commensal microbiota.

# METHODS

## *C. difficile* genome annotation

A new 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 (22). The updated reference genome was annotated using the NCBI Prokaryotic Genome Automatic Annotation Pipeline (53), PATRIC (50), and PROKKA (54) to extract gene features for support of transcriptome pathway enrichment analyses. Bacteriophage loci and genes were identified using PHASTER (55).

## *C. difficile* transcriptional compendium

To generate a transcriptional compendium for *C. difficile*, required for constructing an EGRIN model, a total of 148 publicly available transcriptomes of *C. difficile* 630 were downloaded from the NCBI Gene Expression Omnibus (GEO) repository (56) in March 2020. Downloaded transcriptomes were generated by 11 independent studies (**Table 1**). To integrate this data into a single dataset, we computed the log2 fold-change of each transcriptome with respect to a control condition, as performed in the generation of other transcriptional compendia (57). 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 10 distinct conditional blocks (e.g. sporulation, *fur* deletion), as shown in **Table 1**.

## Construction of the EGRIN model

The EGRIN model for *C. difficile* was constructed in two stages. First, we used cMonkey2 (26), 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 (25), gene promoter sequences from the RSAT database (58), and operon predictions from MicrobesOnline (59) when detecting the 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* motif search was performed using MEME v. 4.12.0 (60). Second, we used the Inferelator (27), 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. We bootstrapped the expression data (20 times) to avoid regression overfitting (27). The Inferelator generates two scores for each TF-module interaction, the corresponding regression coefficient and a confidence score. The second score indicates the likelihood of the interaction. The final set of TF-module interactions was defined as the 704 interactions with the top 10% of highest non-zero confidence scores.

**Experimentally supported literature derived TF regulons**

We mined available literature to compile a list of experimentally supported targets for the 13 partially characterized transcriptional regulators (involved in sporulation, motility, carbon metabolism, among other processes) shown on **Table S1**. The manually compiled regulons represented a total of 1,353 regulatory interactions and involved 1,050 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).

**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 S1**) or functional pathways derived from curated annotation of *C. difficile* genome (22). Only gene modules with adjusted hypergeometric test p-value $\leq 0.05$ and containing four or more genes from the relevant TF regulon or functional pathway were considered enriched.

**Analysis of *in vivo* data**

*In vivo* transcriptomic data from gnotobiotic mice monocolonized with *C. difficile* ATCC43255 or co-colonized with *P. bifermentans* or *C. sardiniense* were analyzed as described (22) using the updated reference genome of ATCC43255 to extract gene features for subsequent analysis with DESEQ2.

**Metabolic model refinement and gene essentiality prediction**

A published genome-scale metabolic model of *C. difficile* 630 strain, icdf834 (41), was used in this study and expanded by adding reactions required for *in vivo* survival of the pathogen. We also curated pathway annotations that were incorrectly designated using default KEGG annotations (61). For example, the TCA, gluconeogenesis and pentose phosphate pathways are incomplete in *C. difficile*. Thus, we updated the annotation of these pathways as part of pyruvate metabolism, butanoate fermentation and Galactose & Tagatose metabolism (**Supplementary File S1**). Initially, 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 764 genes that are predicted in this homology analysis is provided in **Supplementary File S1**. Then, the transcriptome of *C. difficile* profiled from *in vivo* infections of specifically-colonized gnotobiotic mice (22) was mapped onto the icdf834 model using the GIMME algorithm (46). This resulted in a model with 665 active reactions, with no changes in the number of genes. We then expanded the model by including four new genes

554 and 8 new reactions (**Supplementary File S1**) that are required for the growth of the pathogen in
555 the *in vivo* micro-environment, based on KEGG annotations. We named this expanded version of
556 the model as "icdf838". This model represents the *in vivo* state of *C. difficile*. We then applied the
557 constraint-based method for simulating the metabolic steady-state of *C. difficile* using flux-balance
558 analysis (FBA) (46, 62). The initial validation steps involved checking the capacity of the icdf834
559 model to produce biomass in defined media conditions including 1) minimal medium, 2) basal
560 defined medium and 3) complex, nutrient-rich medium (compositions used according to Larocque
561 et al 2014 (42)). Then, we tested the performance of the icdf834 model using gene essentiality
562 predictions by FBA. A gene was considered "essential" if its deletion reduced the biomass by
563 >95%. By this analysis, the model classified each gene as "essential" or "non-essential". We
564 compared the gene essentiality predictions from nutrient-rich media constraints with the available
565 experimental Tn-seq data (35) and deduced the confusion matrix to derive true positive rates
566 (TPR) and false positive rates (FPR). This led to the elucidation of sensitivity and specificity of the
567 model using ROC curve analysis. We then applied the same strategy and predicted the essential
568 genes *in vivo* using FBA with the expanded context-specific network, icdf838. All model
569 simulations related to FBA were performed on MATLAB_R2019a platform using the recent
570 version of COBRA (The COnstraint-Based Reconstruction and Analysis) toolbox (63). *In silico*
571 gene essentiality predictions were performed using the COBRA toolbox 'single-gene-deletion'
572 function in MATLAB. The illustration of essential gene regulatory network *in vivo* was deduced
573 using BioTapsetry tool (http://www.biotapestry.org/).
574
575 ***C. difficile* Web Portal**
576 This portal utilizes the powerful build, search, collaboration, and visualization features of the
577 Drupal content management system. With the two key features of modularity and extensibility,
578 Drupal provides a slim, powerful core that can be readily extended through custom modules and
579 easy-to-use collaborative tools to support information sharing. Based on these key features, we

580 developed this content management system into a data management, analysis, and visualization
581 framework to support *C. difficile* research.
582
583 Due to the complexity of the information provided by the genome and models, it is critical to
584 provide a user-friendly and flexible search and filtering capabilities. By taking advantage of
585 Drupal's built-in search interface and implementing Apache Solr search, we created very powerful
586 search capabilities that will query every information included in the portal database. Moreover,
587 the search interface uses "facets" to allow users to explore a collection of information by applying
588 multiple filters. This combination together with sorting enables users to start with broad searches
589 and then quickly pinpoint specific information.
590
591 In order to provide a comprehensive functional genomics resource for the Cdiff community,
592 genome annotations from several different sources were merged and imported into the Cdiff
593 Portal. Curated genome annotations for *Clostridium difficile* strain 630 published by Monot et al.
594 (21), were downloaded from MicroScope platform (64). Additional functional annotations were
595 downloaded from PATRIC (50) and Uniprot (49) and merged with curated genome annotations.
596 Overall, 4,018 genes were included in the Cdiff Portal. The *C. difficile* genome included 1,030
597 metabolic genes, 309 TFs, 270 sRNAs, 87 tRNAs, 32 rRNAs and 17 miscRNAs. The genome
598 included 1,330 genes with unknown function. Furthermore, gene essentiality data from Dembek
599 et al. (35) was integrated with gene annotations.
600
601

# REFERENCES


1. Monegro AF, Regunath H. 2018. Hospital acquired infectionsStatPearls. StatPearls Publishing.
2. Aktories K. 2011. Bacterial protein toxins that modify host regulatory GTPases. Nat Rev Microbiol 9:487–498.
3. Antunes A, Camiade E, Monot M, Courtois E, Barbut F, Sernova N V, Rodionov DA, Martin-Verstraete I, Dupuy B. 2012. Global transcriptional control by glucose and carbon regulator CcpA in Clostridium difficile. Nucleic Acids Res 40:10701–10718.
4. Saujet L, Monot M, Dupuy B, Soutourina O, 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.
5. Matamouros S, England P, Dupuy B. 2007. Clostridium difficile toxin expression is inhibited by the novel regulator TcdC. Mol Microbiol 64:1274–1288.
6. Mani N, Dupuy B. 2001. Regulation of toxin synthesis in Clostridium difficile by an alternative RNA polymerase sigma factor. Proc Natl Acad Sci 98:5844–5849.
7. Smits WK, Lyras D, Lacy DB, Wilcox MH, Kuijper EJ. 2016. Clostridium difficile infection. Nat Rev Dis Prim 2:1–20.
8. Lyon SA, Hutton ML, Rood JI, Cheung JK, Lyras D. 2016. CdtR regulates TcdA and TcdB production in Clostridium difficile. PLoS Pathog 12:e1005758.
9. Walter BM, Rupnik M, Hodnik V, Anderluh G, Dupuy B, Paulič N, Žgur-Bertok D, Butala M. 2014. The LexA regulated genes of the Clostridium difficile. BMC Microbiol 14:88.
10. Edwards AN, Tamayo R, McBride SM. 2016. A novel regulator controls C lostridium difficile sporulation, motility and toxin production. Mol Microbiol 100:954–971.
11. Martin-Verstraete I, Peltier J, Dupuy B. 2016. The regulatory networks that control Clostridium difficile toxin synthesis. Toxins (Basel) 8:153.
12. Bradshaw WJ, Kirby JM, Roberts AK, Shone CC, Acharya KR. 2017. The molecular structure of the glycoside hydrolase domain of Cwp19 from Clostridium difficile. FEBS J 284:4343–4357.
13. Woods EC, Nawrocki KL, Suárez JM, McBride SM. 2016. The Clostridium difficile Dlt pathway is controlled by the extracytoplasmic function sigma factor $σ$V in response to lysozyme. Infect Immun 84:1902–1916.
14. Neumann-Schaal M, Metzendorf NG, Troitzsch D, Nuss AM, Hofmann JD, Beckstette M, Dersch P, Otto A, Sievers S. 2018. Tracking gene expression and oxidative damage of O2-stressed Clostridioides difficile by a multi-omics approach. Anaerobe 53:94–107.
15. Kint N, Janoir C, Monot M, Hoys S, Soutourina O, Dupuy B, 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.
16. Elena SF, Lenski RE. 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat Rev Genet 4:457–469.
17. Brooks AN, Turkarslan S, Beer KD, Yin Lo F, Baliga NS. 2011. Adaptation of cells to new environments. Wiley Interdiscip Rev Syst Biol Med 3:544–561.
18. McDonald JAK, Mullish BH, Pechlivanis A, Liu Z, Brignardello J, Kao D, Holmes E, Li J V, Clarke TB, Thursz MR, others. 2018. Inhibiting growth of Clostridioides difficile by restoring valerate, produced by the intestinal microbiota. Gastroenterology 155:1495–1507.
19. Vemuri RC, Gundamaraju R, Shinde T, 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.
20. Riedel T, Bunk B, Thürmer A, Spröer C, Brzuszkiewicz E, Abt B, Gronow S, Liesegang H,

652 Daniel R, Overmann J. 2015. Genome resequencing of the virulent and multidrug-
653 resistant reference strain Clostridium difficile 630. Genome Announc 3:e00276--15.
654 21. Monot M, Boursaux-Eude C, Thibonnier M, Vallenet D, Moszer I, Medigue C, Martin-
655 Verstraete I, Dupuy B. 2011. Reannotation of the genome sequence of Clostridium
656 difficile strain 630.
657 22. Girinathan BP, DiBenedetto N, Worley JN, Peltier J, Lavin Ri, Delaney ML, Cummins C,
658 Onderdonk AB, Gerber GK, Dupuy B, others. 2020. The mechanisms of in vivo
659 commensal control of Clostridioides difficile virulence. bioRxiv.
660 23. Bouillaut L, Self WT, Sonenshein AL. 2013. Proline-dependent regulation of Clostridium
661 difficile Stickland metabolism. J Bacteriol 195:844–854.
662 24. Bouillaut L, Dubois T, Francis MB, Daou N, Monot M, Sorg JA, Sonenshein AL, Dupuy B.
663 2019. Role of the global regulator Rex in control of NAD+-regeneration in Clostridioides
664 (Clostridium) difficile. Mol Microbiol 111:1671–1688.
665 25. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva
666 NT, Roth A, Bork P, others. 2016. The STRING database in 2017: quality-controlled
667 protein--protein association networks, made broadly accessible. Nucleic Acids Res
668 gkw937.
669 26. Reiss DJ, Plaisier CL, Wu W-J, Baliga NS. 2015. cMonkey2: Automated, systematic,
670 integrated detection of co-regulated gene modules for any organism. Nucleic Acids Res
671 43:e87.
672 27. Arrieta-Ortiz ML, Hafemeister C, Bate AR, Chu T, Greenfield A, Shuster B, Barry SN,
673 Gallitto M, Liu B, Kacmarczyk T, Santoriello F, Chen J, Rodrigues CD, Sato T, Rudner
674 DZ, Driks A, Bonneau R, Eichenberger P. 2015. An experimentally supported model of
675 the Bacillus subtilis global transcriptional regulatory network. Mol Syst Biol 11.
676 28. Peterson EJR, Reiss DJ, Turkarslan S, Minch KJ, Rustad T, Plaisier CL, Longabaugh
677 WJR, Sherman DR, Baliga NS. 2014. A high-resolution network model for global gene
678 regulation in Mycobacterium tuberculosis. Nucleic Acids Res 42:11291–11303.
679 29. Brooks AN, Reiss DJ, Allard A, Wu W-J, Salvanha DM, Plaisier CL, Chandrasekaran S,
680 Pan M, Kaur A, Baliga NS. 2014. A system-level model for the microbial regulatory
681 genome. Mol Syst Biol 10:740–740.
682 30. Saujet L, Pereira FC, Serrano M, Soutourina O, Monot M, Shelyakin P V, Gelfand MS,
683 Dupuy B, Henriques AO, Martin-Verstraete I. 2013. Genome-wide analysis of cell type-
684 specific gene transcription during spore formation in Clostridium difficile. PLoS Genet
685 9:e1003756.
686 31. Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, Thorsson V, Shannon P,
687 Johnson MH, Bare JC, Longabaugh W, Vuthoori M, Whitehead K, Madar A, Suzuki L,
688 Mori T, Chang D-E, Diruggiero J, Johnson CH, Hood L, Baliga NS. 2007. A predictive
689 model for transcriptional control of physiology in a free living cell. Cell 131:1354–65.
690 32. Dineen SS, Villapakkam AC, Nordman JT, Sonenshein AL. 2007. Repression of
691 Clostridium difficile toxin gene expression by CodY. Mol Microbiol 66:206–219.
692 33. Underwood S, Guan S, Vijayasubhash V, Baines SD, Graham L, Lewis RJ, Wilcox MH,
693 Stephenson K. 2009. Characterization of the sporulation initiation pathway of Clostridium
694 difficile and its role in toxin production. J Bacteriol 191:7296–7305.
695 34. Wolfe CJ, Kohane IS, Butte AJ. 2005. Systematic survey reveals general applicability of"
696 guilt-by-association" within gene coexpression networks. BMC Bioinformatics 6:227.
697 35. Dembek M, Barquist L, Boinett CJ, Cain AK, Mayho M, Lawley TD, Fairweather NF,
698 Fagan RP. 2015. High-throughput analysis of gene essentiality and sporulation in
699 Clostridium difficile. MBio 6:e02383--14.
700 36. Bradshaw WJ, Bruxelle J-F, Kovacs-Simon A, Harmer NJ, Janoir C, Péchiné S, Acharya
701 KR, Michell SL. 2019. Molecular features of lipoprotein CD0873: A potential vaccine
702 against the human pathogen Clostridioides difficile. J Biol Chem 294:15850–15861.

37. Steglich M, Hofmann JD, Helmecke J, Sikorski J, Spröer C, Riedel T, Bunk B, Overmann J, Neumann-Schaal M, 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.
38. Nawrocki KL, Wetzel D, Jones JB, Woods EC, McBride SM. 2018. Ethanolamine is a valuable nutrient source that impacts Clostridium difficile pathogenesis. Environ Microbiol 20:1419–1435.
39. Dingle TC, Mulvey GL, Armstrong GD. 2011. Mutagenic analysis of the Clostridium difficile flagellar proteins, FliC and FliD, and their contribution to virulence in hamsters. Infect Immun 79:4061–4067.
40. Lee ASY, Song KP. 2005. LuxS/autoinducer-2 quorum sensing molecule regulates transcriptional virulence gene expression in Clostridium difficile. Biochem Biophys Res Commun 335:659–666.
41. Kashaf SS, Angione C, 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.
42. Larocque M, Chénard T, Najmanovich R. 2014. A curated C. difficile strain 630 metabolic network: prediction of essential targets and inhibitors. BMC Syst Biol 8:117.
43. Theriot CM, Koenigsknecht MJ, Carlson Jr PE, Hatton GE, Nelson AM, Li B, Huffnagle GB, Li JZ, Young VB. 2014. Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection. Nat Commun 5:3114.
44. Janoir C, Denève C, Bouttier S, Barbut F, Hoys S, Caleechum L, Chapetón-Montes D, Pereira FC, Henriques AO, Collignon A, others. 2013. Adaptive strategies and pathogenesis of Clostridium difficile from in vivo transcriptomics. Infect Immun 81:3757–3769.
45. Karasawa T, Ikoma S, Yamakawa K, Nakamura S. 1995. A defined growth medium for Clostridium difficile. Microbiology 141:371–375.
46. Becker SA, Palsson BO. 2008. Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol 4:e1000082.
47. Jenior ML, Leslie JL, Young VB, Schloss PD. 2017. Clostridium difficile colonizes alternative nutrient niches during infection across distinct murine gut microbiomes. Msystems 2.
48. Galperin MY, Makarova KS, Wolf YI, Koonin E V. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res 43:D261--D269.
49. Consortium TU. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res 45:D158--D169.
50. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, others. 2014. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res 42:D581--D591.
51. Gößner AS, Picardal F, Tanner RS, Drake HL. 2008. Carbon metabolism of the moderately acid-tolerant acetogen Clostridium drakei isolated from peat. FEMS Microbiol Lett 287:236–242.
52. Sakai S, Inokuma K, Nakashimada Y, Nishio N. 2008. Degradation of glyoxylate and glycolate with ATP synthesis by a thermophilic anaerobic bacterium, Moorella sp. strain HUC22-1. Appl Environ Microbiol 74:1447–1452.
53. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 44:6614–6624.
54. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

55. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res 44:W16--W21.
56. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, others. 2012. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41:D991--D995.
57. Moretto M, Sonego P, Dierckxsens N, Brilli M, Bianco L, Ledezma-Tejeida D, Gama-Castro S, Galardini M, Romualdi C, Laukens K, Collado-Vides J, Meysman P, Engelen K. 2016. COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. Nucleic Acids Res 44:D620-3.
58. Nguyen NTT, Contreras-Moreira B, Castro-Mondragon JA, Santana-Garcia W, Ossio R, Robles-Espinoza CD, Bahin M, Collombet S, Vincens P, Thieffry D, others. 2018. RSAT 2018: regulatory sequence analysis tools 20th anniversary. Nucleic Acids Res 46:W209--W214.
59. Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, Friedland GD, Huang KH, Keller K, Novichkov PS, Dubchak IL, Alm EJ, Arkin AP. 2010. MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic Acids Res 38:D396-400.
60. Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME suite. Nucleic Acids Res 43:W39--W49.
61. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45:D353--D361.
62. Orth JD, Thiele I, Palsson BØ. 2010. What is flux balance analysis? Nat Biotechnol 28:245–248.
63. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, Haraldsdóttir HS, Wachowiak J, Keating SM, Vlasov V, others. 2019. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v. 3.0. Nat Protoc 14:639–702.
64. Vallenet D, Calteau A, Cruveiller S, Gachet M, Lajus A, Josso A, Mercier J, Renaux A, Rollin J, Rouy Z, others. 2017. MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. Nucleic Acids Res 45:D517--D528.
65. Dineen SS, McBride SM, Sonenshein AL. 2010. Integration of metabolism and virulence by Clostridium difficile CodY. J Bacteriol 192:5350–5362.
66. Soutourina O, Dubois T, Monot M, Shelyakin P V, Saujet L, Boudry P, Gelfand MS, Dupuy B, 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.
67. Antunes A, Martin-Verstraete I, Dupuy B. 2011. CcpA-mediated repression of Clostridium difficile toxin gene expression. Mol Microbiol 79:882–899.
68. Dubois T, Dancer-Thibonnier M, Monot M, Hamiot A, Bouillaut L, Soutourina O, Martin-Verstraete I, Dupuy B. 2016. Control of Clostridium difficile physiopathology in response to cysteine availability. Infect Immun 84:2389–2405.
69. Berges M, Michel A-M, Lassek C, Nuss AM, Beckstette M, Dersch P, Riedel K, Sievers S, Becher D, Otto A, others. 2018. Iron regulation in Clostridioides difficile. Front Microbiol 9:3183.
70. El Meouche I, Peltier J, Monot M, Soutourina O, Pestel-Caron M, Dupuy B, 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.
71. Fimlaid KA, Bond JP, Schutz KC, Putnam EE, Leung JM, Lawley TD, Shen A. 2013. Global analysis of the sporulation pathway of Clostridium difficile. PLoS Genet 9.

# FIGURES

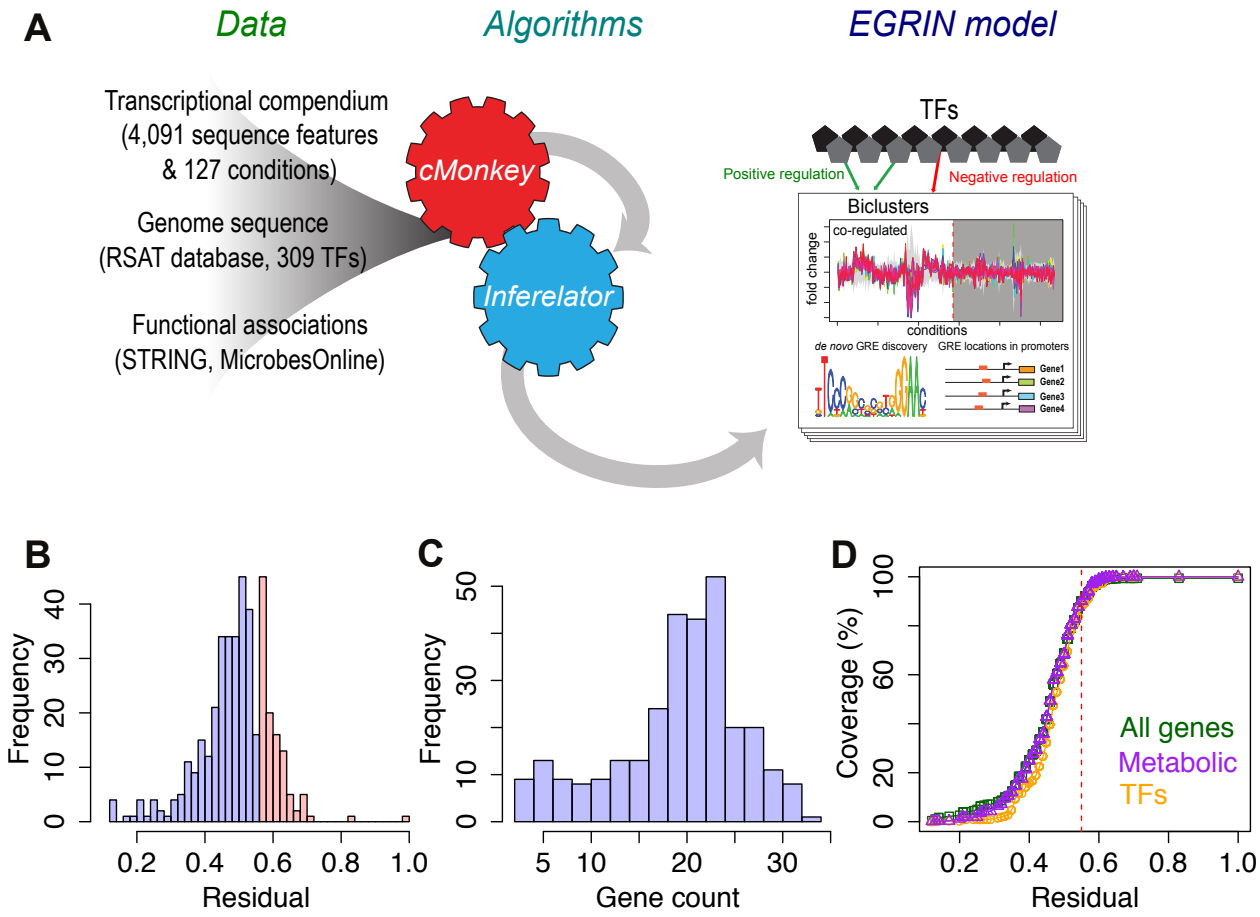
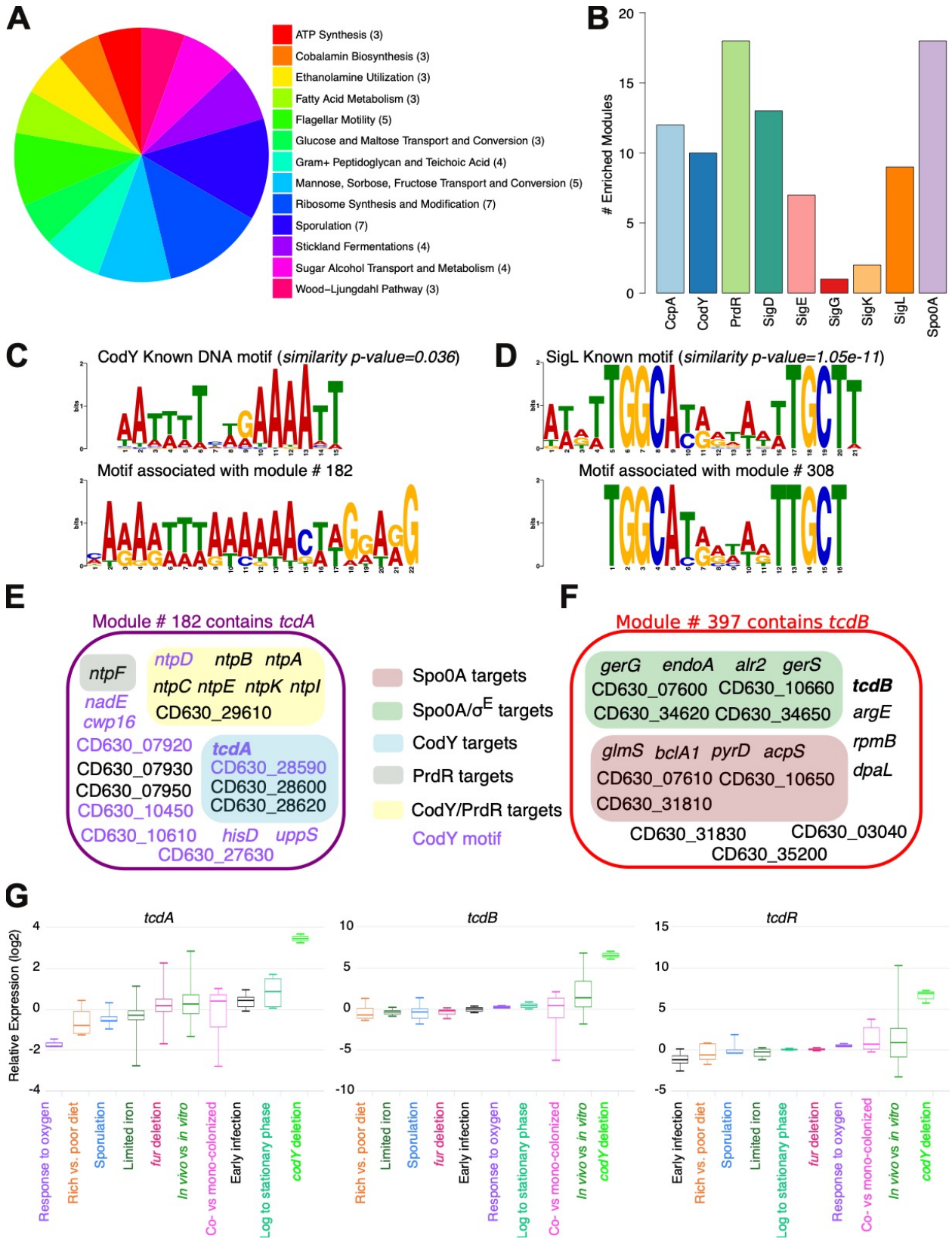
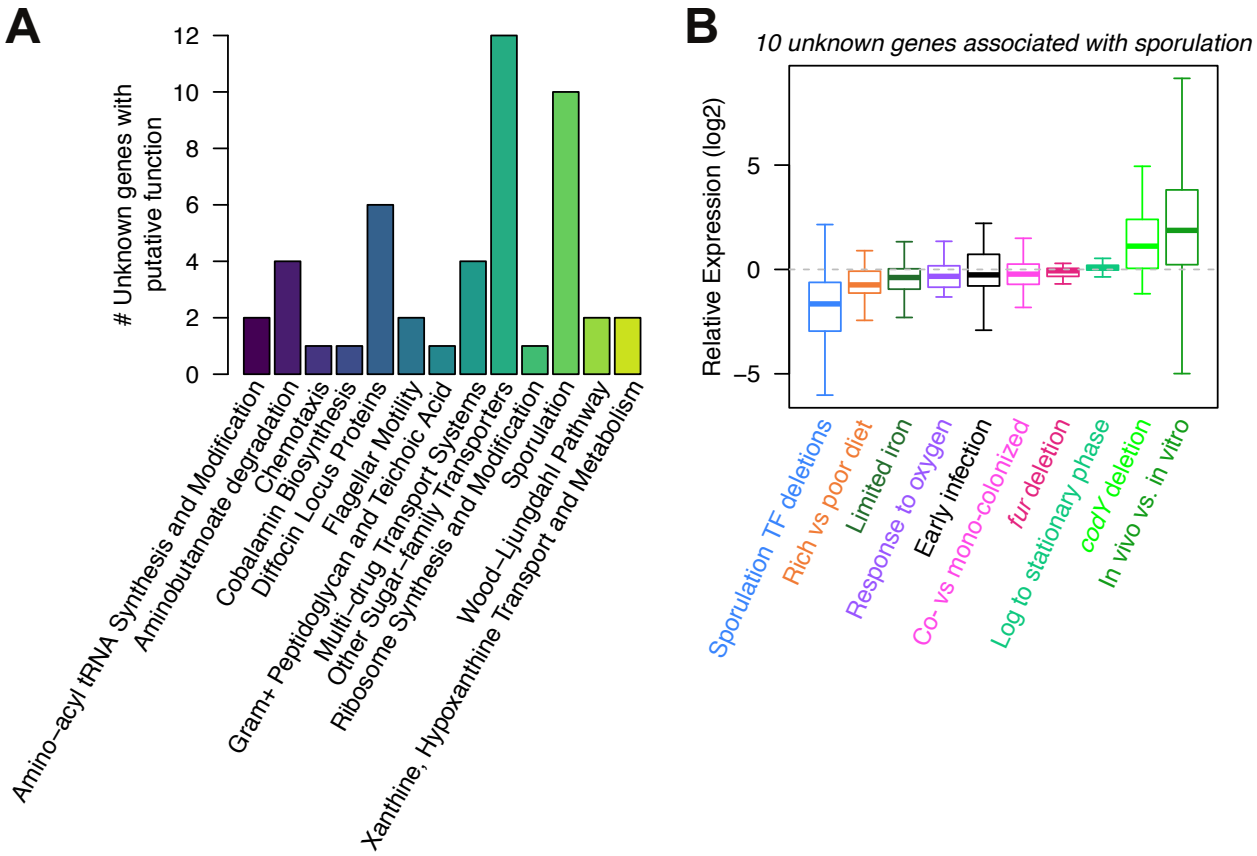



**Figure 1. Inference pipeline and general properties of the resulting Environment Gene Regulatory Influence Network (EGRIN) model of *C. difficile*.** (A) Framework used to build the EGRIN model. (B) Distribution of residual values for the 406 detected co-regulated gene modules. 297 gene modules had residual values equal or lower than 0.55 (shown in purple) and were labelled 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.



817 **Figure 2. The Environment Gene Regulatory Influence Network model of *C. difficile***
818 **recapitulates known biology of the pathogen.** (A) Co-regulated gene modules are enriched
819 with functional terms derived from expert curated annotation of the *C. difficile* genome (22). The
820 pie chart shows terms over-represented in three or more modules. Number of modules associated
821 with each functional term is shown in parenthesis. (B) Enriched gene modules among nine (out
822 of 13) manually-defined and experimentally supported TF regulons (compiled from publicly
823 available data in Table S1). (C) EGRIN identified the known DNA binding motif of CodY (65). (D)
824 EGRIN also identified the known DNA binding motif of SigL (66). Motif comparisons were
825 performed using Tomtom (60). (E) The EGRIN model recapitulated the previously reported
826 influence of CodY on *tcdA* expression. The module #182 contains *tcdA*, it is enriched with
827 members of the CodY regulon and contains a motif (shown in panel C) similar to the
828 experimentally determined CodY motif. (F) The EGRIN model also captured the interaction
829 between toxin expression and sporulation via module #397 that contains *tcdB* and is enriched
830 with genes regulated by sporulation-related transcriptional regulators. **(**G) Expression profiles of
831 *tcdAB* and *tcdR* (positive regulator of the pathogenicity loci). Highest expression of the toxin genes
832 and their activator was observed in the *codY* single deletion condition (light green box).
833

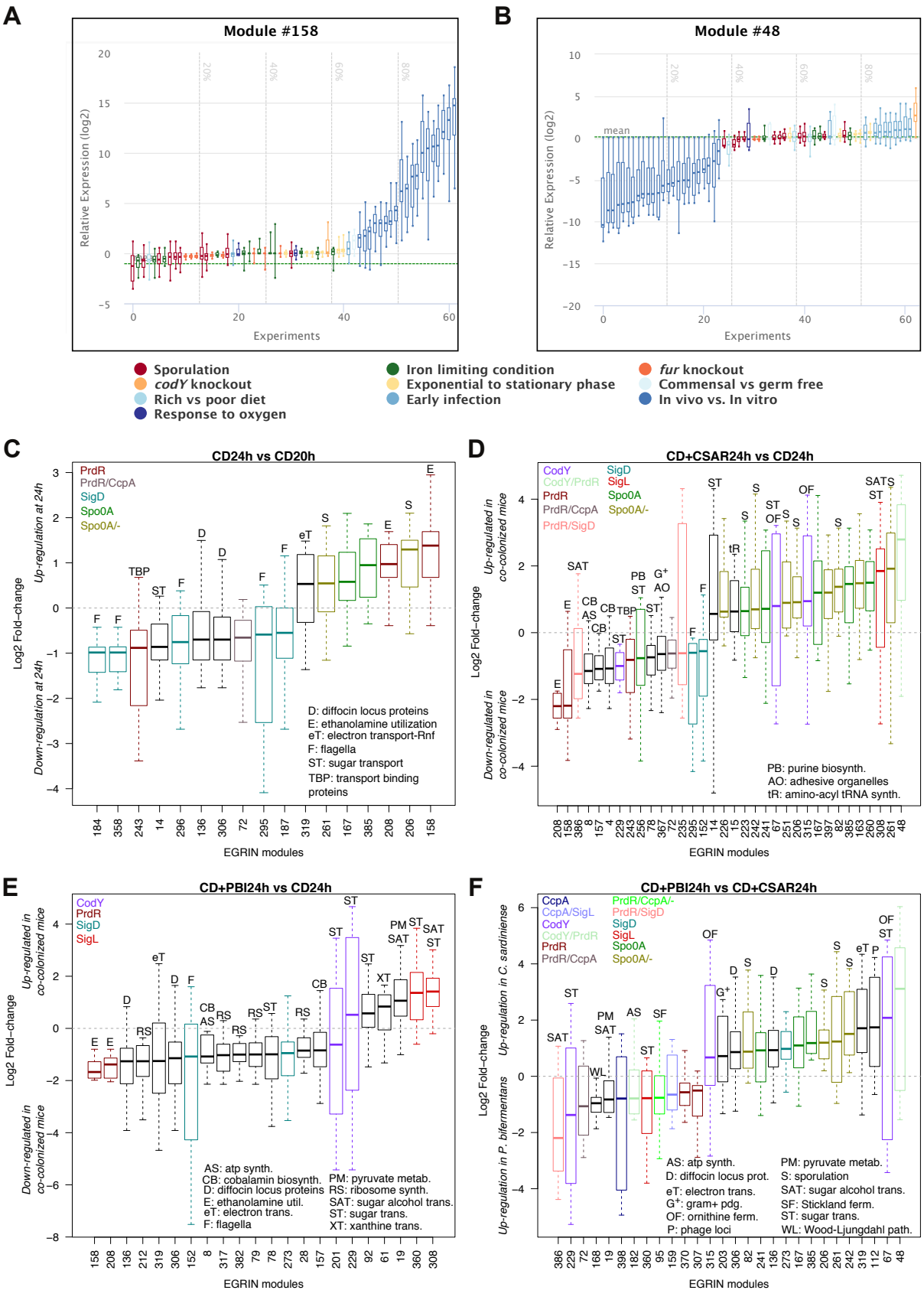



C

| Predicted role | Enriched modules | Enriched TF regulons | Uncharacterized genes |
|---|---|---|---|
| Other Sugar-family Transporters | 43 | NA | CD630_04371, CD630_09341, CD630_12100, CD630_13011, CD630_13781, CD630_13782, CD630_15970, CD630_17720, CD630_20101, CD630_20751, CD630_29661, CD630_31110 |
| Sporulation | 206, 251 | Spo0A, SigE, SigG | CD630_01310, CD630_02140, CD630_10650, CD630_10660, CD630_19280, CD630_23740, CD630_23750, CD630_24430, CD630_26340, CD630_32590 |
| Diffocin Locus Proteins | 136 | NA | CD630_12420, CD630_12430, CD630_12440, CD630_13760, CD630_14150, CD630_14160 |
| Aminobutanoate degradation | 48 | CodY, PrdR | CD630_05100 , CD630_05101, CD630_08760, CD630_08780 |
| Multi-drug transport systems | 29 | NA | CD630_03790, CD630_03800, CD630_10991, CD630_33681 |

**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 co-regulated gene modules significantly enriched with

839 specific functional terms. (A) Barplot with the number of unknown genes associated with each
840 functional term (from the *C. difficile* genome annotation in Girinathan et al. (22) ). (B) The
841 involvement of 10 uncharacterized genes in sporulation was supported by the observed strongest
842 and significant down-regulation in single deletion strains of transcriptional regulators of
843 sporulation (*spo0A*, *sigEFGK*, *spoIIID*). (C) Locus tag of uncharacterized genes associated with
844 selected functional terms.
845




846

**Figure 4. The EGRIN model identifies TFs driving the *in vivo* response of *C. difficile* when interacting with gut commensals *P. bifermentans* (PBI) and *C. sardiniense* (CSAR).** (A) Expression profile of module #158. (B) Expression profile of module #48. (C) EGRIN modules enriched with genes differentially expressed (absolute log2 fold-change > 1 and adjusted p-value < 0.05) in *C. difficile* mono-colonized mice at 24 vs 20 hours of infection. X-axis shows module IDs. Modules were annotated according to their functional enrichment and overlap with manually curated TF regulons (Table S1). (D) Enriched EGRIN modules in *C. sardiniense*+*C. difficile* co-colonized mice vs *C. difficile* mono-colonized mice at 24 hours of infection. Due to space constraint, only abbreviations of functional terms not shown in other panels are displayed. (E) Enriched EGRIN modules in *P. bifermentans*+*C. difficile* co-colonized mice vs *C. difficile* mono-colonized mice at 24 hours of infection. (F) Enriched EGRIN modules in *P. bifermentans*+*C. difficile* co-colonized mice vs *C. sardiniense*+*C. difficile* co-colonized mice at 24 hours of infection. For all comparisons, only modules with absolute median fold-changes > 0.5, and enriched with TF regulons or functional categories are displayed.

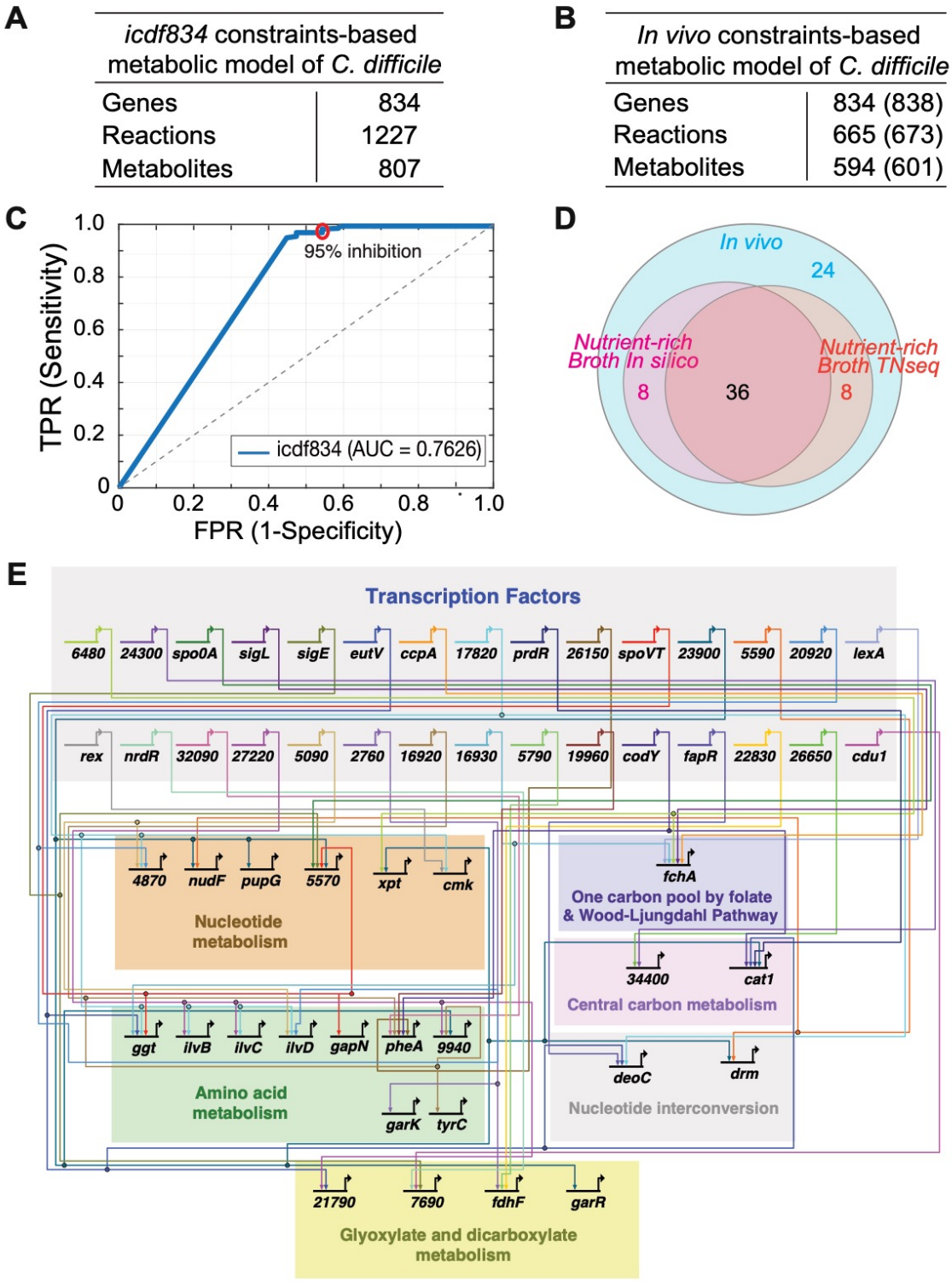
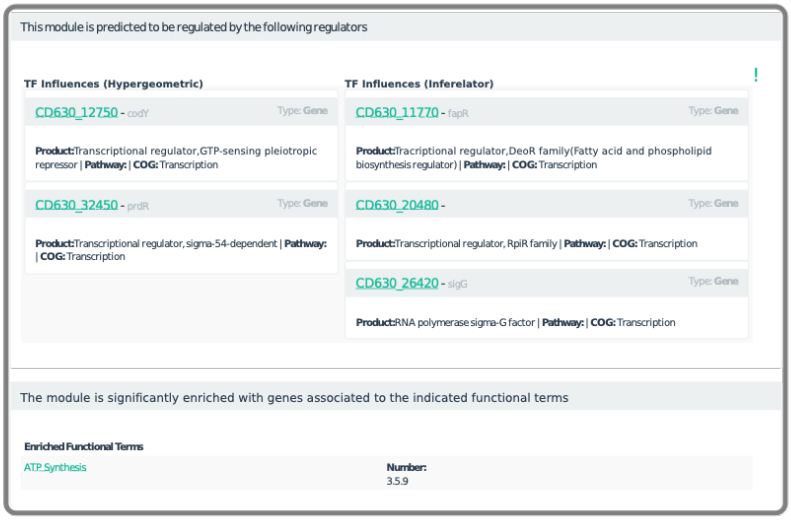



**Figure 5. Metabolic model predictions.** (A) Details of the *in vitro* metabolic model of *C. difficile* 630 (41). (B) Details of icdf838 metabolic model of *C. difficile* 630. Initial *in vivo* model was derived

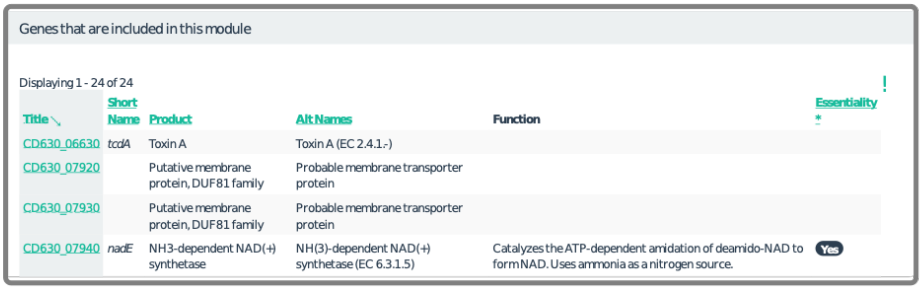
using the GIMME algorithm (46) where only the active reactions are included from *in vivo* transcriptome. The numbers in parentheses indicate the number of genes, reactions and metabolites in the icdf838 model after adding the required *in vivo* exchanges and transports. (C) ROC curve showing the accuracy of icdf834-predicted gene essentiality in nutrient rich medium evaluated against a Tn-seq functional screen (35). Red circle indicates the 95% growth inhibition as threshold. (D) Venn diagram showing the number of model-predicted essential genes for growth of *C. difficile 630 in vitro* vs *in vivo*. (E) BioTapestry visualization of *in vivo* gene regulatory network for *C. difficile* 630: All 24 *in vivo*-specific essential genes that are regulated by transcription factors (TFs) are shown. Transcriptional regulators are derived from EGRIN. The network includes all TFs that regulate more than four *in vivo* essential genes. The genes and TFs shown as five digit numbers represent the nomenclature preceded by 'CD630_' (e.g. 27220 indicates CD630_27220).

A


B




C




881

882 **Figure 6. An example module page of the *C. difficile* Portal.** Module #182, associated with
883 CodY and shown in Fig 2D is used as an example. (A) Each module page includes general
884 statistics of the module (residual score, gene count), displays the module expression profile in the
885 compiled transcriptional compendium and the detected motifs. (B) A module page also offers
886 information about the potential transcriptional regulators of the module. Putative regulators are
887 defined based on over-representation of manually compiled TF regulons (assessed using
888 hypergeometric test) and based on the Inferelator predictions. (C) Each module page includes a
889 list of its gene members with a brief description of each gene. This information includes gene
890 name, product, alternative names, function and essentiality. In the example, only the first four
891 genes (out of 24) are shown. The user can click in any gene to visit the corresponding gene page.
892
893

# TABLES

**Table 1. Datasets used for generating *C. difficile* transcriptional compendium**

| Condition | GEO Series Accession | # Transcriptomes[a] | # Controls[b] |
|---|---|---|---|
| Early infections (0h, 30 mins, 60 mins, 120 mins) | GSE18407 | 12 | NA[c] |
| *In vivo* vs *in vitro* (8h, 14h, 38h) | GSE43305 | 32 | NA[c] |
| Iron limitation | GSE109453 GSE120189 | 15 | 15[d] |
| *fur* deletion | GSE69218 GSE120189 | 12 | 12[d] |
| Response to oxygen | GSE109175 | 3 | 3 |
| Response to commensals and diet | GSE60751 | 8 | 8[e] |
| Rich diet vs Poor diet | GSE60751 | 8 | 8[e] |
| Transition from exponential to stationary phase | GSE115054 | 16 | NA[c] |
| *codY* KO | GSE23192 | 3 | NA[c] |
| Sporulation (*spo0A* KO, *sigEFGK* KOs, *spoIIID* KO) | GSE45977 GSE63777 | 18 | 6 |

[a]Refers to the number of arrays used as numerator when estimating log2 ratios

[b]Control arrays were averaged and used as denominator when estimating log2 ratios

[c]Not Applicable. Dual channel array and therefore the control was included in each array

[d]Six samples used as controls were also considered as main transcriptome in other comparisons

[e]Samples used as controls were also considered as main transcriptome in other comparisons

# SUPPLEMENTARY FIGURES

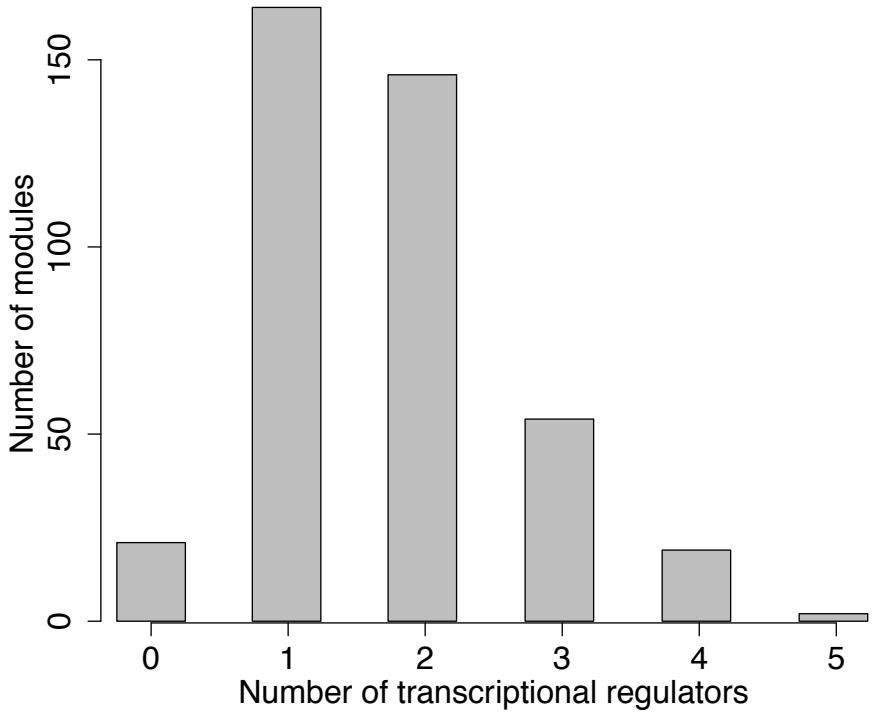
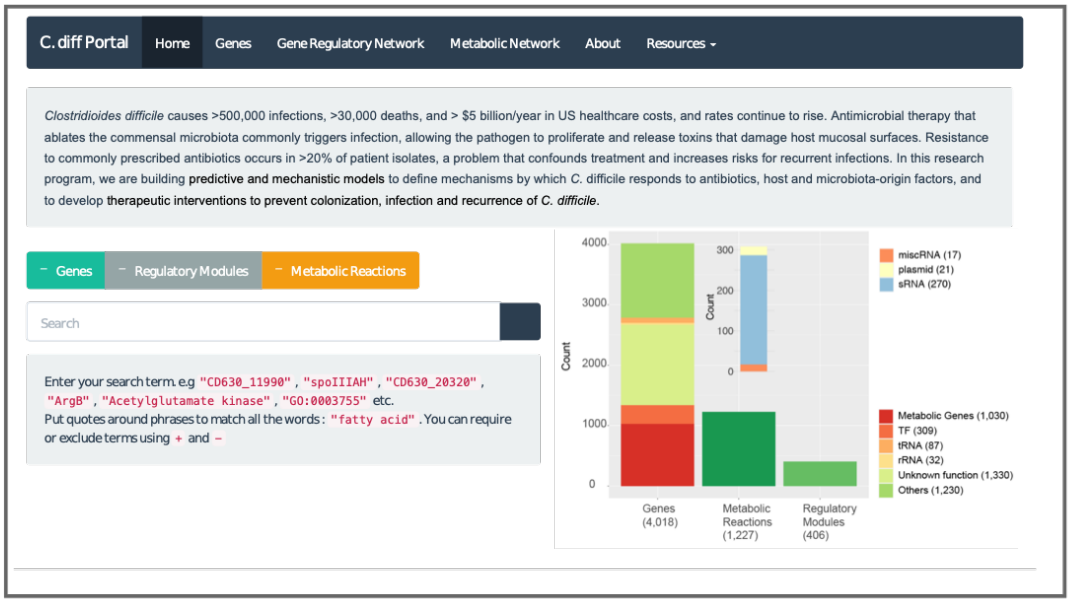



**Figure S1. Number of Inferelator-predicted transcriptional regulators of modules in the EGRIN model.**

A




B


909

910 **Figure S2. The Cdiff Web Portal.** (A) Home page of the Cdiff Portal. Users can explore the
911 gene regulatory network model and the metabolic model for *C. difficile*. In addition, all files are
912 accessible in the resource tab. The search bar facilitates website exploration. (B) An example
913 gene page of the *C. difficile* Portal.

914
915

# SUPPLEMENTARY TABLES

**Table S1. Compiled TF regulons**

| Regulator | Regulon Size | Supporting Data | Reference |
|---|---|---|---|
| CcpA | 194[a] | Protein-DNA binding, transcriptomics, *in silico*[b] | (3, 67) |
| CodY | 160[c] | Protein-DNA binding, transcriptomics | (65) |
| Fur | 19 | Transcriptomics, *in silico*[b] | (68, 69) |
| PrdR | 181 | Transcriptomics | (24) |
| SigB | 57 | Transcriptomics | (15) |
| SigD | 159 | Transcriptomics | (70) |
| SigE | 96 | Transcriptomics | (30) |
| SigF | 25 | | |
| SigG | 46 | | |
| SigH | 40 | Transcriptomics, *in silico*[b] | (4) |
| SigK | 54 | Transcriptomics | (30) |
| SigL | 46[d] | Transcriptomics, *in silico*[b] | (66) |
| Spo0A | 276[e] | Transcriptomics | (71) |

[a]Only genes classified as CcpA-dependent (in the presence or absence of glucose) by Antunes et al. 2012 (3) were included

[b]*In silico* search of the binding motif of the corresponding regulator within the genes and their promoter sequences

[c]Only genes negatively influenced by CodY were included

[d]Only genes with SigL motif and down-regulated in *sigL* deletion were included

[e]Only genes positively influenced by Spo0A were included

928 **Table S2. General properties of modules associated with the pathogenicity loci**

| Gene | Locus tag | Module[a] | Residual | Functional enrichment[b] | Enriched TF regulons[b] | Inferelator predicted regulators[b] |
|---|---|---|---|---|---|---|
| *tcdR* | CD_06590 | 31 | 0.46 | NF | NF | SigG, SpoVT |
| | | 336 | 0.54 | NF | NF | FapR, CD_16930 |
| *tcdB* | CD_06600 | 284 | 0.63 | NF | NF | CD_06930, CD_17820 |
| | | 397 | 0.49 | NF | Spo0A, SigE | SpoVT, CD_06290 |
| *tcdE* | CD_06610 | 31 | 0.46 | NF | NF | SigG, SpoVT |
| | | 138 | 0.61 | NF | NF | CD_06290, CD_12390, SpoVT |
| *tcdA* | CD_06630 | 182 | 0.56 | ATP synthesis | CodY, PrdR | FapR, CD_20480, SigG |
| | | 264 | 0.71 | NF | NF | SigV |
| *tcdC* | CD_06640 | 146 | 0.52 | NF | NF | CD_27320, PhoU |
| | | 395 | 0.51 | NF | NF | CD_18100, CD_35440 |

929 [a]Modules that contain the corresponding member of the pathogenicity loci

930 [b]Refers to the information shown for the corresponding module on the Cdiff Web Portal

931

932 **Table S3. Model predicted essential genes *in vivo***

| Locus tag | Gene name | Gene product/function | Pathway | In silico predictions | | Expt |
|---|---|---|---|---|---|---|
| | | | | Essentiality in vivo | Essentiality nutrient-rich in vitro | Essentiality Nutrient-rich in vitro (Tn-seq) |
| CD630_09940 | G12WB-1109 | serine-pyruvate aminotransferase | Alanine, aspartate and glutamate metabolism | Essential | Non-Essential | Non-Essential |
| CD630_05800 | gapN | glyceraldehyde-3-phosphate dehydrogenase | Valine, leucine and isoleucine metabolism | Essential | Non-Essential | Non-Essential |
| CD630_15340 | ggt | gamma-glutamyltranspeptidase | Alanine, aspartate and glutamate metabolism | Essential | Non-Essential | Non-Essential |
| CD630_23430 | cat1 | succinyl-CoA:coenzyme A transferase | Butanoate fermentation; Methionine Biosynthesis | Essential | Non-Essential | Non-Essential |
| CD630_34400 | G12WB-3619 | glycoside hydrolase-type carbohydrate-binding protein | Glycolysis | Essential | Non-Essential | Non-Essential |
| CD630_30840 | garK | glycerate kinase | Glycine, serine and threonine metabolism | Essential | Non-Essential | Non-Essential |
| CD630_28130 | garR | tartronate semialdehyde reductase | Glyoxylate and dicarboxylate metabolism | Essential | Non-Essential | Non-Essential |
| CD630_33170 | fdhF | formate dehydrogenase-H | Glyoxylate and dicarboxylate metabolism | Essential | Non-Essential | Non-Essential |
| CD630_21790 | G12WB-2336 | anaerobic dehydrogenase | Glyoxylate and dicarboxylate metabolism | Essential | Non-Essential | Non-Essential |
| CD630_07690 | G12WB-880 | oxidoreductase subunit | Glyoxylate and dicarboxylate metabolism | Essential | Non-Essential | Non-Essential |

| | | | | | | |
|---|---|---|---|---|---|---|
| CD630_12240 | pupG | purine nucleoside phosphorylase | Purine metabolism | Essential | Non-Essential | Non-Essential |
| CD630_07190 | fchA | methenyltetrahydrofolate cyclohydrolase | One carbon pool by folate; Wood-Ljungdahl pathway | Essential | Non-Essential | Non-Essential |
| CD630_15660 | ilvB | acetolactate synthase large subunit | Valine, leucine and isoleucine metabolism | Essential | Non-Essential | Non-Essential |
| CD630_12230 | drm | phosphopentomutase | Nucleotide interconversion | Essential | Non-Essential | Non-Essential |
| CD630_20140 | ilvD | dihydroxy-acid dehydratase | Valine, leucine and isoleucine metabolism | Essential | Non-Essential | Non-Essential |
| CD630_15020 | deoC | deoxyribose-phosphate aldolase | Nucleotide interconversion | Essential | Non-Essential | Non-Essential |
| CD630_18390 | tyrC | prephenate dehydrogenase | Phenylalanine, tyrosine and tryptophan biosynthesis | Essential | Non-Essential | Non-Essential |
| CD630_18360 | pheA | bifunctional chorismate mutase/prephenate dehydratase | Phenylalanine, tyrosine and tryptophan biosynthesis | Essential | Non-Essential | Non-Essential |
| CD630_12200 | nudF | NUDIX family hydrolase | Purine metabolism | Essential | Non-Essential | Non-Essential |
| CD630_23300 | xpt | xanthine phosphoribosyltransferase | Purine metabolism | Essential | Non-Essential | Non-Essential |
| CD630_05570 | G12WB-669 | uridine kinase | Pyrimidine metabolism | Essential | Non-Essential | Non-Essential |
| CD630_04870 | G12WB-599 | carbon-nitrogen hydrolase | Pyrimidine metabolism | Essential | Non-Essential | Non-Essential |
| CD630_18160 | cmk | cytidylate kinase | Pyrimidine metabolism | Essential | Non-Essential | Non-Essential |

| | | | | | | |
|---|---|---|---|---|---|---|
| CD630_15650 | ilvC | ketol-acid reductoisomerase | Valine, leucine and isoleucine metabolism | Essential | Non-Essential | Non-Essential |