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 Pathogenesis of Bacterial Anaerobes, Institut Pasteur, Université de Paris, UMR-CNRS2001, France

*Lead Contact: nitin.baliga@isbscience.org

SUMMARY

Though *Clostridioides difficile* is among the most studied anaerobes, the interplay of metabolism and regulation that underlies its ability to colonize the human gut is unknown. We have compiled public resources into three models and a portal to support comprehensive systems analysis of *C. difficile*. First, by leveraging 151 transcriptomes from 11 studies we generated a regulatory model (EGRIN) that organizes 90% of *C. difficile* genes into 297 high quality conditional co-regulation modules. EGRIN predictions, validated with independent datasets, recapitulated and extended regulons of key transcription factors, implicating new genes for sporulation, carbohydrate transport and metabolism. Second, by advancing a metabolic model, we discovered that 15 amino acids, diverse carbohydrates, and 10 metabolic genes are essential for *C. difficile* growth within an intestinal environment. Finally, by integrating EGRIN with the metabolic model, we developed a PRIME model that revealed unprecedented insights into combinatorial control of essential processes for *in vivo* colonization of *C. difficile* and its interactions with commensals. We have developed an interactive web portal (http://networks.systemsbiology.net/cdiff-portal/) to disseminate all data, algorithms, and models to support collaborative systems analyses of *C. difficile*.

32 # INTRODUCTION

33 *Clostridioides difficile*, the etiology of pseudomembranous colitis, causes more than 500,000
34 infections, 30,000 deaths, and $5 billion per year in US healthcare costs (Monegro and Regunath,
35 2018). Infections arise through a variety of conditions that modulate the pathogen's ability to
36 colonize and expand in the gut. Antibiotic ablation of the commensal microbiota creates altered
37 nutrient states in intestinal environments due to lack of competition for nutrients from host, dietary
38 or microbial origin. The pathogen modifies its endogenous metabolism to respond to these altered
39 states, which stimulates subsequent cellular programs that can promote enhanced colonization
40 and growth. Stress and starvation conditions within *C. difficile* populations trigger responses that
41 lead to sporulation, biofilm formation and release of mucosal damaging toxins (Aktories, 2011;
42 Antunes et al., 2012; Saujet et al., 2011).

43 Symptomatic infection requires the production of toxins from the *C. difficile* pathogenicity locus
44 (PaLoc), which includes the genes *tcdA, tcdB* and *tcdE* that respectively encode the A and B
45 toxins and holin involved in toxin export (Govind and Dupuy, 2012). The PaLoc also contains
46 *tcdR*, a sigma factor specific for the toxin gene promoters, and *tcdC*, a putative TcdR anti-sigma
47 factor (Carter et al., 2011; Cartman et al., 2012; Mani and Dupuy, 2001; Matamouros et al., 2007).
48 *C. difficile* elaborates toxin under starvation conditions to extract nutrients from the host and
49 promote spore shedding (Edwards et al., 2016; Martin-Verstraete et al., 2016; Walter et al., 2014).
50 Regulation of PaLoc expression occurs via a complex network of transcription factors (TFs) and
51 small molecule inputs, of which direct primary regulators have been well described, but more
52 complex and combinatorial effects remain unclear (Martin-Verstraete et al., 2016). Toxin
53 production further triggers rapid and profound host immune responses, including release of
54 reactive oxygen species which substantially alters the redox state of the gut environment, and
55 other innate immune responses that can induce *C. difficile* stress responses to cell wall, oxidative,

56 and other damaging stimuli (Bradshaw et al., 2017; Kint et al., 2017; Neumann-Schaal et al.,
57 2018; Woods et al., 2016). As per all microbes, *C. difficile* adapts to complex, dynamic
58 environments through changes in metabolism coordinated by a gene regulatory network (Brooks
59 et al., 2011; Elena and Lenski, 2003). However, the mechanisms by which the gene regulatory
60 network and metabolic pathways integrate to modulate *C. difficile* pathogenesis remain ill-defined
61 (McDonald et al., 2018; Vemuri et al., 2017).

62 The *C. difficile* 630 (CD630) genome encodes 4,018 genes, with ~309 candidate TFs (including
63 sigma factors), 1,030 metabolic genes, and 1,330 genes (>30%) with unknown function (Monot
64 et al., 2011; Riedel et al., 2015). The clinical ATCC43255 strain of *C. difficile,* used to capitulate
65 symptomatic infections in mouse models, encodes 4,117 genes and ~327 putative TFs, of which
66 ~97% are orthologous to genes encoded in the CD630 strain (Girinathan et al., 2021). To address
67 questions regarding the broader systems-level interplay among genes in colonization and
68 infection, we used computational modeling and network inference algorithms to construct an
69 Environment and Gene Regulatory Influence Network (EGRIN) model for *C. difficile.* This model
70 leverages a compendium of 151 published transcriptomes that surveyed responses of CD630 in
71 diverse contexts. The EGRIN model consists of modules of putatively co-regulated genes
72 identified based on their co-expression over subsets of conditions, enrichment of functional
73 associations, chromosomal proximity, and presence of cis-acting gene regulatory elements
74 (GREs) within their promoter regions indicating regulation by the same TFs. Further, using
75 regression analysis, EGRIN also captures the combinatorial regulation of genes within each
76 module as a function of the weighted influences of TFs. The model supports a systems-level
77 understanding of the infective capacity of this obligate anaerobe under different *in vitro* and *in vivo*
78 conditions.

79 In addition to EGRIN, we have advanced a metabolic network model of *C. difficile* to understand
80 how conditional regulation manifests physiologically, by adding reactions and associated genes

supporting the exchange of nutrients required for growth in intestinal environments. Integration of transcriptional and metabolic networks into a **P**henotype of **R**egulatory influences **I**ntegrated with **M**etabolism and **E**nvironment (**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 *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* (Bouillaut et al., 2013, 2019). The PRIME model successfully identified enhanced epistasis between *ccpA* and *codY* in the presence of a protective gut commensal species. 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/.

98 # RESULTS

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

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

116

117 The quality of modules within the EGRIN model was evaluated using residual scores, which reflect
118 the coherence of gene co-expression patterns. The lower the residual score, the higher the quality
119 of the module. We determined using an empirical approach that a residual cutoff of 0.55 identified
120 a functionally meaningful set of 297 high quality modules (73% of the total 406 modules) based
121 on the relative enrichment of related functions within modules that passed filtering (**Fig. 1B**). This
122 empirically determined threshold was similar to the threshold used to identify high quality EGRIN

123 gene modules for *Mycobacterium tuberculosis* (Peterson et al., 2014). Altogether, the 297 high
124 quality modules captured transcriptional regulation of 3,617 genes (90%) in CD630, with average
125 membership of 20 genes per module (**Fig. 1C-D**). These metrics were consistent with models
126 developed for other organisms (Brooks et al., 2014; Peterson et al., 2014), a remarkable finding
127 given that the transcriptional dataset used to construct the *C. difficile* model was less than 10%
128 the size of ones used to construct models for other species.
129

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

132 We tested the accuracy of the EGRIN model to reconstruct previously characterized regulons and
133 recapitulate key aspects of *C. difficile* biology. To do so, we performed gene enrichment analysis
134 within modules using an updated annotation of *C. difficile* genome (Girinathan et al., 2021). This
135 analysis identified 93 of 297 modules (31%) with significant enrichment of genes with related
136 functions in 45 pathways (hypergeometric test adjusted p-value ≤ 0.05). Among these pathways,
137 14 were over-represented in three or more modules (**Fig. 2A**), demonstrating the capacity of the
138 model to discover conditional partitioning of cellular processes. We also investigated whether the
139 EGRIN model had identified known regulatory interactions between TFs and their target genes.
140 We compiled from literature the regulons (i.e., target genes) of 13 previously characterized TFs
141 in *C. difficile*, representing a network of 1,349 TF-gene interactions (**Table S2**). Notably, a total
142 of 57 modules (19% of all high-quality modules) were significantly enriched with nine of these TF
143 regulons (**Fig. 2B**). The EGRIN model recapitulated 514 of the 1,208 (42.5%) previously
144 characterized interactions. This value is consistent with the recall rate of the EGRIN model for *M.*
145 *tuberculosis* (41%-49%) (Peterson et al., 2014). The poor recall of the remaining four regulons
146 (141 regulatory interactions) could be due to underrepresentation of gene expression data from
147 relevant conditions in which these regulons are conditionally active. This analysis also uncovered
148 combinatorial regulation of genes across 19 modules (i.e., enriched with more than one TF

149 regulon). Consistent with the known hierarchical scheme for regulation of sporulation (Saujet et
150 al., 2013), expression of 161 genes across at least eight modules were putatively influenced by
151 Spo0A in combination with one or more alternative sigma factors implicated in sporulation (e.g.,
152 SigE). EGRIN also predicted CcpA contributions in six additional modules in combination with
153 CodY, PrdR and SigL, illustrating the complexity of modular transcriptional regulation in *C. difficile*.
154 The EGRIN modules also detected co-regulation of genes within and across functionally related
155 operons. For example, module #152, which is enriched with the SigD regulon, contains 16 genes
156 that are part of four operons including the flagellar operon *flgG1G-fliMN*-CD630_02720-*htpG,* in
157 addition to *pyrBKDE*, CD630_30270-CD630_30280-*malY*-CD630_30300, and CD630_32430-
158 *prdA* (**Fig. S2A**).
159
160 The clustering of genes in the EGRIN model is constrained by the *de novo* discovery of conserved
161 GRE(s) within their promoters in order to cluster genes that are co-regulated, and not just co-
162 expressed. The GREs represent putative binding sites for TFs that are often independently
163 implicated by the Inferelator and protein-DNA interaction maps as regulators of genes within the
164 same module (Bonneau et al., 2007; Brooks et al., 2014; Peterson et al., 2014). For instance, the
165 regression-based analysis by the Inferelator assigned 138 TFs as putative regulators of the 297
166 gene modules, hypothesizing putative TF-GRE associations. While the paucity of characterized
167 binding sites in *C. difficile* limited the degree to which these TF-GRE mappings could be validated,
168 the predicted context-specific TF regulation of genes within modules can serve as a guide for
169 performing ChIP-seq experiments in relevant conditions to iteratively improve the mechanistic
170 accuracy of the EGRIN model. Notably, we determined that the GREs within promoters of genes
171 in modules #182 and #308 recapitulated the known consensus binding sequences for CodY and
172 SigL, which are predicted regulators of those modules, and are among the very few TFs in *C.*
173 *difficile* for which binding sites have been characterized (**Fig. 2C-D and Fig. S2B-C**)(Dineen et
174 al., 2007; Soutourina et al., 2020).

175

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

177 We evaluated capacity for the EGRIN model to recall known mechanisms of PaLoc regulation,
178 and to provide new information regarding complex regulatory and small molecule effects on toxin
179 production. The EGRIN model captured certain previously described effects of CodY on toxin
180 gene expression (**Fig. 2E**), as shown in module #182, which is enriched with members of the
181 CodY regulon including *tcdA*. In agreement with EGRIN-predicted CodY regulation of PaLoc
182 genes in module #182, genes encoding the toxin *tcdA* and its regulator *tcdR* were significantly
183 overexpressed upon deletion of *codY* (**Fig. 2G**). Interestingly, *tcdB* (which is not part of module
184 #182) was also significantly upregulated in the *codY* deletion strain, suggesting that this effect
185 might be an indirect consequence of disrupted CodY regulation of *tcdR* (**Fig. 2G**). It has been
186 assumed that CodY acts on PaLoc gene expression primarily through its repression of *tcdR.*
187 Putative lower affinity binding sites have been suggested in the toxin gene promoter regions
188 (Dineen et al., 2007). The presence of the CodY motif (**Fig. 2C**) in most members of module #182,
189 including *tcdA* (purple font in **Fig. 2E**) suggests direct influence of CodY on *tcdA* gene expression.
190 The EGRIN model also identified previously reported connections between sporulation and toxin
191 production (Underwood et al., 2009). The *tcdB* toxin gene was assigned to module #397, which
192 was significantly enriched with genes controlled by Spo0A, the master regulator of sporulation
193 (**Fig. 2F**). Additional members of the PaLoc were assigned to other modules, supporting the
194 presence of multiple condition-dependent promoters within the PaLoc (**Table S3**).

195

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

197 Approximately 33% of gene features in the CD630 genome have unknown functions. Thus, the
198 *C. difficile* EGRIN model emerges as a resource to assign putative functions to uncharacterized
199 genes based on functional associations among co-regulated genes (i.e., guilt-by-
200 association)(Wolfe et al., 2005). We predicted putative functions for 48 uncharacterized genes by

201 mining underlying functional enrichment of modules under different experimental conditions
202 (**Table S4**; *see methods*). These 48 previously uncharacterized genes were associated with 13
203 functional categories, including “Sporulation” and “Other sugar-family transporters” (**Fig. 3A**).
204
205 Ten genes were putatively assigned sporulation-related functions based on their co-regulation in
206 the sporulation associated modules #206 and #251 (**Fig. 3B)**. Module #206 includes seven stage
207 III sporulation genes (*spoIIIAA*, *spoIIIAB*, *spoIIIAC*, *spoIIIAD*, *spoIIIAF*, *spoIIIAG* and *spoIIIAH*),
208 and two stage IV sporulation genes (*spoIV*, *spoIVA*) (**Fig. S3A**). Module #251 includes the
209 sporulation-associated alternative sigma factors SigG and SigE (located in the same operon) (**Fig.**
210 **S3B**). Decreased expression of the 10 putative sporulation genes upon deletion of sporulation-
211 associated sigma factors suggested putative roles within the mother cell or the forespore. Seven
212 genes are likely associated with mother cell-specific roles based on their decreased expression
213 in *sigE* (six genes) and *sigK* (one gene) deletion strains (**Table S2**). Two additional genes were
214 downregulated in a *sigG* deletion strain, suggesting putative functions in the forespore. Notably,
215 Tn-seq studies for gene essentiality in *C. difficile* identified seven of these 10 genes as required
216 for sporulation (Dembek et al., 2015) (**Fig. S3A-B**).
217
218 Module #48 contains two adjacent operons (*4hbd-cat2*-CD630_23400-*abfD* and *sucD-cat1*)
219 associated with aminobutanoate degradation (**Fig. S3C**). Both operons are regulated by CodY
220 and PrdR. Hence, we predicted that the four uncharacterized genes in this module may be also
221 involved in amino acid metabolism (**Table S4**). In support of this hypothesis, CD630_08760 and
222 CD630_08780 are both differentially expressed upon *codY* deletion. Recent studies also suggest
223 that CD630_ 08760 may function as a tyrosine transporter per its homology to the CodY-regulated
224 neighbor gene, CD630_08730 (Bradshaw et al., 2019). A finding supported by the observed
225 decreases in tyrosine uptake and Stickland fermentation in clinical isolates lacking CD630_08760
226 and CD630_08780 (Steglich et al., 2018). Altogether, we assigned 48 genes to putative functions

227 in sporulation (10 genes), sugar transport (12 genes), among others. These examples illustrate
228 how integration of multiple sources of evidence for context-specific co-regulation of genes in
229 EGRIN can also predict putative functions for uncharacterized genes. EGRIN can also aid in
230 designing experiments (e.g., perturbations of TFs and target genes) to validate predicted function
231 assignments in relevant environmental context.
232
233 **EGRIN uncovers differentially active regulatory networks during *in vivo* infection**
234 We investigated the differential expression of EGRIN modules across multiple *in vivo* studies
235 (Janoir et al., 2013; Janvilisri et al., 2010) to discover the regulatory mechanisms that drive *C.*
236 *difficile's* colonization and adaption to *in vivo* environments. Analyses identified 680 genes across
237 43 modules that were significantly upregulated *in vivo*, while 1,325 genes across 82 modules
238 were significantly downregulated. Notably, three modules were also differentially regulated during
239 early stages of infection (**Supplementary File S1**). For example, this analysis discovered the *in*
240 *vivo* activation and repression of module #158 and module #48, respectively (**Fig. 4A-B**). Notably,
241 module #48 was upregulated during early infection and it is enriched with members of the CodY
242 and PrdR regulons, as described above. Module #158 is enriched for putative PrdR and EutV co-
243 regulated ethanolamine utilization genes, including *eut* operons for a two-component histidine
244 kinase sensing system and carboxysome structural proteins that house the ethanolamine
245 fermentative enzymes (Nawrocki et al., 2018) (**Fig. S3D**). Ethanolamine is prevalent within gut
246 secretions and is also released from damaged host tissues, providing a readily available carbon
247 and nitrogen source for *C. difficile*. The predicted co-regulation of this gene module by PrdR
248 (Pearson correlation = -0.29 and P-value = 9.7e-04) suggests additional *in vivo* functions of this
249 regulator to optimize *C. difficile's* metabolism in gut environments. The negative relation between
250 *prdR* expression and module #158 expression indicates that enhanced utilization of ethanolamine
251 as it becomes a newly available carbon and nitrogen source *in vivo* may contribute to the

252 decreased survival observed in hamsters infected with a *prdR* deletion strain (Bouillaut et al.,
253 2019).
254
255 With capacity to identify intestinal contributions to *C. difficile* responses, we leveraged the EGRIN
256 model to analyze commensal modulation of the pathogen's virulence, using transcriptomic
257 datasets from gnotobiotic mice that were mono-colonized with the mouse-infective strain *C.*
258 *difficile* ATCC43255 or co-colonized with *C. difficile* and the protective gut commensal species
259 *Paraclostridium bifermentans* (PBI), or infection-worsening species *Clostridium sardiniense*
260 (CSAR) (Girinathan et al., 2021). These datasets were not used in model construction. By
261 mapping sets of differentially expressed genes into the EGRIN model we uncovered modules
262 across 18 cellular processes and their associated TFs that were differentially regulated in the
263 presence of PBI or CSAR (**Fig. 4C-F**).
264
265 One Spo0A-enriched module (module #385) was upregulated by 24h of infection in mono-
266 colonized mice (**Fig. 4C**). The same module was upregulated by 24h of infection in CSAR co-
267 colonized mice, in addition to three other modules enriched with sporulation genes (modules #82,
268 #206, #261 in **Fig. 4D**). These three modules were also enriched with the Spo0A regulon. On the
269 other hand, no sporulation-enriched modules were detected by 24h of infection in PBI co-
270 colonized mice (**Fig. 4E**). Comparison of CSAR co-colonized mice and PBI co-colonized mice
271 discovered four sporulation-enriched modules (modules #82, #206, #242 and #261) with
272 increased expression in the virulent context (i.e., presence of CSAR) (**Fig. 4F**). These findings
273 were confirmed with the high levels of spore release in expanded populations of vegetative *C.*
274 *difficile* when co-colonized with CSAR (Girinathan et al., 2021). Overall, this analysis suggested
275 that the sporulation pathway is an indicator of *C. difficile* disease, reinforcing the Spo0A-mediated
276 link between sporulation and toxin production recapitulated by the model (**Fig. 2F**).
277

278 Module #319 contains multiple genes associated with electron transport via Rnf ferredoxin
279 systems, and steps in glycolytic, butanoate and succinate metabolic pathways (**Fig. S3E**). This
280 module was consistently downregulated in mice co-colonized with the protective commensal PBI
281 (**Fig. 4E**), and was upregulated in CSAR co-colonized mice when compared to PBI co-colonized
282 mice (**Fig. 4F**). These findings highlight activation of multiple co-regulated energy generating
283 pathways in hypervirulent states of *C. difficile* which may contribute to virulence and to the
284 pathogen's responses to changing redox states brought on by severe host inflammatory
285 responses. Remarkably, the EGRIN model identified the NAD+/NADH sensing regulator Rex as
286 a potential repressor of module #319 (Inferelator regression coefficient = -0.075). The observed
287 downregulation of module #319 in PBI co-colonized mice indicates increased Rex activity when
288 *C. difficile* was introduced with this commensal and associated nutrient depleted state to support
289 *C. difficile* growth (Girinathan et al., 2021). In support of a positive influence of PBI on Rex activity,
290 we observed downregulation (suggesting decreased activity) and upregulation of *rex* mRNA level
291 in the compiled transcriptional compendium during *in vitro* growth and early infection, respectively
292 (**Fig. S3F**).
293
294 Four modules enriched with the SigD-regulated genes encoding subunits of flagella (modules
295 #184, #295, #296 and #358) were downregulated in mono-colonized mice at 24h (**Fig. 4C**),
296 indicating repression of motility to divert cellular resources toward pathogenesis. This finding is
297 supported by increased virulence of *C. difficile* strains lacking a functional flagella (Dingle et al.,
298 2011). In summary, the analysis of differential regulation patterns of EGRIN modules has
299 generated new mechanistic insights and hypotheses for the interplay of specific genes and TFs
300 for sporulation, energy production, and flagella synthesis that might underlie the enhanced and
301 subdued virulence of *C. difficile* in different contexts.
302
303 **Metabolic network analyses elucidate *in vivo* metabolic adaptations of *C. difficile***

304 To investigate how specific genes within *C. difficile* contribute to *in vivo* phenotypes needed to
305 develop symptomatic infection we extended a previously developed icdf834 metabolic model for
306 CD630 (Kashaf et al., 2017; Larocque et al., 2014). We added four genes (CD630_08700,
307 CD630_08680, CD630_17090 and CD630_10810) that encode reactions for molybdenum
308 utilization and cofactor synthesis, and added exchange reactions to account for *C. difficile's*
309 capacity to utilize mannitol, fructose, sorbitol, raffinose, succinate and butanoate (Janoir et al.,
310 2013; Theriot et al., 2014). In total, we added 46 reactions. We refer to this updated model as
311 icdf838 (**Fig. 5A**). Lastly, the model derived from CD630 was compared with the gene feature
312 content from *C. difficile* ATCC43255, used as a model organism to investigate symptomatic
313 infections in mice. The two strains shared 92% of metabolic genes and predicted pathways (768
314 out of 838 genes in the icdf838 model have homology with the ATCC43255 strain;
315 **Supplementary File S2**).
316
317 We validated the completeness and accuracy of this model by confirming its ability to predict
318 biomass production in three different *in vitro* media compositions: 1) minimal medium, 2) basal
319 defined medium and 3) complex, nutrient-rich medium (see Larocque et al., 2014 for media
320 compositions). The model accurately predicted *C. difficile's* requirements for six amino acids:
321 cysteine, leucine, isoleucine, proline, tryptophan and valine (Karasawa et al., 1995). Using a Tn-
322 seq generated genome-wide fitness screen (Dembek et al., 2015), we tested the accuracy of
323 model-predicted importance of each gene in supporting the growth of *C. difficile* in nutrient rich
324 conditions using a receiver-operator characteristic (ROC) curve. The area under the ROC curves
325 for the icdf834 and icdf838 models were 0.763 (p-value = 0.015) and 0.764 (p-value =0.029),
326 respectively (**Fig. 5B**), indicating that both models significantly outperformed a random model.
327
328 We next extended and applied the model to predict *C. difficile* behaviors and gene essentiality *in*
329 *vivo. C. difficile* transcriptomes from specifically-colonized gnotobiotic mice (Girinathan et al.,

330 2021) were used as input into the GIMME algorithm (Becker and Palsson, 2008). Per the *in vivo*
331 gene expression, we determined that 712 of the 1,273 total reactions in the icdf838 model were
332 active and likely to be important for colonization and growth over the course of symptomatic
333 infection (**Fig. 5C**). Leveraging information from the *in vitro* studies, the model made two notable
334 predictions *in vivo* regarding the pathogen's metabolism. First, the icdf838 model predicted 15
335 amino acids to be required for *C. difficile* growth *in vivo*, in contrast to the 6 required *in vitro*
336 (**Supplementary File S2**). These amino acids included the dominant Stickland-fermented amino
337 acids that were also required *in vitro,* including proline and branched chain amino acids, and
338 additional amino acids including arginine, glutamate, lysine and methionine, which also have
339 multiple cellular functions in cell wall synthesis, nitrogen cycling, and responses to oxidative
340 stress. Secondly, the model predicted *C. difficile's* switch from preferential use of glucose as a
341 carbon source *in vitro* in complex media, to simultaneous utilization *in vivo* of diverse carbohydrate
342 sources including fructose, galactose, maltose, and sugar alcohols such as mannitol and sorbitol,
343 to promote colonization and growth (**Supplementary File S2**). Seven of these carbohydrate
344 sources were described in other *in vivo* mouse infection studies illustrating support for these
345 findings across *C. difficile* strains, and in germfree and conventional mouse models
346 (**Supplementary File S2**) (Janoir et al., 2013; Jenior et al., 2017; Theriot et al., 2014).
347

348 **Integration of EGRIN and metabolic model reveals how the interplay of regulation and**
349 **metabolism governs *C. difficile* adaptations to *in vivo* host environments.**

350 To evaluate how transcriptional regulation modulates *C. difficile*'s metabolic and physiological
351 responses, we integrated the transcriptional and metabolic networks into a **P**henotype of
352 **R**egulatory influences **I**ntegrated with **M**etabolism and **E**nvironment (**PRIME**) model (Immanuel et
353 al., 2021). The PRIME model evaluates context-specific TF essentiality, an approach used to
354 uncover vulnerabilities of *M. tuberculosis* (Immanuel et al., 2021). To generate a PRIME model
355 for *C. difficile*, we first inferred a transcriptional network of 1,405 TF-gene interactions involving

356 215 TFs and 733 metabolic genes. This step was necessary to quantify combinatorial and
357 conditional influence of TFs in metabolic genes and their associated metabolic fluxes, and to
358 compensate for the poor coverage and lack of information about TF-gene interactions in *C.*
359 *difficile*. Condition-specific PRIME models were generated by integrating *C. difficile* transcriptomic
360 data from mice mono- and co-colonized (*C. difficile* and PBI), and using metabolic constraints for
361 *in vitro* growth.
362
363 PRIME predicted essential metabolic genes and networks that promote *C. difficile's* growth *in vivo*
364 (i.e., in mono-colonized mice). In simulating the consequences of single gene deletions, PRIME
365 predicted that *in vivo* growth of the pathogen in mono-colonized mouse would decrease by ≥65%
366 with individual knockouts of 10 genes involved in one carbon-cycling reactions, and nucleotide
367 and central carbohydrate metabolism (**Fig. 5D, Fig. 6A and Table S5**). These metabolic
368 pathways represent new potential targets that drive *C. difficile's* colonization and subsequent
369 growth, factors required to develop symptomatic infections. Model predictions also illustrated *C.*
370 *difficile's* predicted shift from carbohydrate utilization towards amino acid utilizing pathways *in*
371 *vivo*, as shown by the enhanced set of 15 amino acids, including the preferred Stickland donor
372 and acceptor amino acids (leucine and proline) known to support metabolism and growth (Janoir
373 et al., 2013; Jenior et al., 2017; Theriot et al., 2014). Notably, many of these amino acids show
374 high abundance within the gut lumen in gnotobiotic and in antibiotic-treated conventional mice
375 that enhance *C. difficile's* capacity to colonize and expand (Girinathan et al., 2021).
376
377 One of the key advantages of PRIME is its extended capability to evaluate TF essentiality. We
378 evaluated the performance of PRIME by generating ROC curves as described above. The area
379 under the ROC curve for PRIME gene essentiality predictions was comparable with
380 corresponding values for the icdf834 and icdf838 models (**Fig. S4**). Given the influence of CcpA
381 and CodY on toxin production and regulation of multiple metabolic genes (Antunes et al., 2012;

382 Dineen et al., 2010), we leveraged PRIME to predict the phenotypic consequences of single and
383 double knockouts in *ccpA* and *codY* in both mono- and PBI co-colonized mouse infections (**Table**
384 **1**). Using condition-specific models, PRIME predicted that three genes regulated by CcpA (*pgsA*,
385 *ribC* and CD630_02440) were essential in both mono- and co-colonized conditions (**Fig. 6A**).
386 Similarly, PRIME predicted that three genes regulated by CodY (*drm*, *glmM*, *guaA*) were essential
387 in mono- and co-colonized mouse and one additional gene (*pupG*) was only essential in PBI co-
388 colonized mouse (**Fig. 6A**). Strikingly, PRIME predicted that a double knockout of CcpA and CodY
389 would have a synergistic consequence on inhibiting *C. difficile* biomass (as compared to the
390 effects of individual knockouts in each TF), *especially* in PBI co-colonized mouse. Five out of the
391 six PRIME essentiality predictions were validated by performing infections of germ free and PBI
392 colonized mouse with wild type and mutant strains of *C. difficile* (**Table 1**). In addition, PRIME
393 offered mechanistic insights into the conditional phenotypes of *C. difficile* by revealing how the
394 single and double knockouts of *ccpA* and *codY* remodel the metabolic state of the pathogen in
395 mono- and co-colonized contexts (**Fig. 6B**). For example, reactions associated with the one
396 carbon pool by folate pathway were inactive in the co-colonized condition but were differentially
397 affected by the gene knockouts in the mono-colonized mouse. PRIME also predicted that flux for
398 reactions associated with teichoic acid biosynthesis and terpenoid backbone biosynthesis was
399 invariable in the mono-colonized mice but sensitive to gene deletions in co-colonized mice. In
400 summary, by providing mechanistic insights into how transcriptional regulation propagates
401 throughout the metabolic network to manifest in fitness, PRIME offers a new capability to
402 investigate how *C. difficile* adapts to complex biotic and abiotic changes within the *in vivo* gut
403 environment.
404

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

406 We have released a new *C. difficile* Web Portal (http://networks.systemsbiology.net/cdiff-portal/)
407 to provide a discovery and collaboration gateway for the *C. difficile* scientific community. The

408 portal aims to accelerate the advancement of the science and understanding of *C. difficile* biology,
409 and the relationship between gene regulation, metabolism and virulence. Within the portal users
410 can access publicly available datasets (e.g., transcriptional compendia), models, software and
411 supporting resources. The Portal includes information on more than 4,000 *C. difficile* genes, 1,273
412 metabolic reactions, and 406 co-regulated gene modules. Notably, the EGRIN model can be
413 interactively explored, and queried with gene set enrichment analysis to elucidate regulatory
414 networks that may be relevant for a user-specified group of genes. Genes can be explored in the
415 context of genome annotations, expression profiles, regulatory and metabolic membership, and
416 other functional genomic information across different databases including COG, Uniprot, and
417 PATRIC (Consortium, 2017; Galperin et al., 2015; Wattam et al., 2014). The portal provides
418 access to detailed information on (1) Genes, (2) predicted Gene Modules, and (3) Metabolic
419 Reactions (**Fig. S5A**).
420
421 Each gene module page includes summary statistics for the module, context-specific differential
422 expression patterns of module genes, GREs, TF regulatory influences, enrichment of biological
423 functions and pathways, and information on each regulon member gene. The module pages are
424 structured to facilitate the assessment of the quality and statistical significance of the modules
425 and highlight functional connections, while allowing users to implement their own filters (e.g.,
426 regression coefficient and adjusted p-value thresholds) (**Fig. S5B**). The portal includes a table of
427 metabolic reactions with details of each reaction, associated genes, metabolites, and sub-
428 systems. Metabolites and sub-systems are defined as taxonomic vocabularies that collect and
429 group associated reactions to identify related metabolic processes. In addition, the portal provides
430 access to algorithms, software, and data, and will include information about animal models,
431 strains, and other *C. difficile* relevant community resources. As additional datasets are
432 communicated, model predictions and tools will be successively enhanced to support systems-

433 level analyses and assist in hypothesis generation in *C. difficile* biology and to enable tangible
434 clinical interventions.

435 # DISCUSSION

436 The obligate anaerobe *C. difficile* is unique among gut anaerobes in possessing a diverse carbon
437 source metabolism to enable colonization and growth in gut environments. These systems further
438 exist within a complex network of gene regulatory modules that modulate growth, energy balance,
439 and stress responses *in vivo*. Capacity to understand these systems-level integration points has
440 remained challenging in the absence of robust systems biology models to infer *C. difficile's in vivo*
441 behaviors. We acknowledge the detailed studies from multiple groups over prior decades that
442 provided a critical mass of information on *C. difficile's* nutrient and gene-level responses to
443 support development of an EGRIN model, the first for a gut anaerobe and toxigenic species. We
444 emphasize that this information, the most for any obligate anaerobe, still represents a small
445 fraction of that normally used to develop thorough EGRIN models. Recent improvements in the
446 genetic manipulation of *C. difficile*, including the mouse infective strain ATCC43255 (Peltier et al.,
447 2020), open new capacity to probe the GREs modulating critical aspects of its metabolism, growth
448 and virulence, from a systems-level perspective.
449
450 The *C. difficile* EGRIN model enables a number of predictions relevant to *in vivo* disease. For
451 example, the PrdR regulator of the pathogen's Stickland proline reductase (*prdABDEF*) and other
452 genes, has long been hypothesized to have a role in PaLoc gene expression through as-yet
453 unknown mechanisms. EGRIN predictions included gene module #182, which identified
454 combined PrdR and CodY effects on *tcdA* gene expression (suggested by the enrichment of
455 module #182 with both PrdR and CodY regulons), providing a regulatory integration point and
456 broader set of co-regulated genes to support further experimental analyses of co-regulation
457 between these two TFs, including effects on PaLoc expression. Biclustering also identified
458 interactions between Spo0A, another regulator hypothesized to modulate PaLoc expression, and
459 *tcdB* expression in module #397. The identified modules, associated genes and regulators

460 provide new information to support further experimental investigation of combinatorial effects of
461 these and other regulators identified in PaLoc gene-associated modules. The EGRIN model also
462 predicted PrdR as a critical regulator *in vivo* through its systems-level effects on critical metabolic
463 and regulatory networks supporting the pathogen's colonization, metabolism and growth. These
464 interactions mediated by PrdR involve multiple direct and indirect effects upon other modules
465 (e.g., module #48 and module #158) and aspects of the pathogen's metabolism and gene
466 regulation.
467
468 The current model did not identify all experimentally known regulators of PaLoc expression,
469 including SigD regulation of TcdR, and effects of other more recently identified PaLoc regulators
470 such as RstA and LexA, for which limited datasets exist from wild type or targeted deletion
471 mutants under relevant nutritional and other environmental perturbations. Additional datasets with
472 isogenic regulator mutants will likely improve predictions while further validating previously
473 defined biologic effects. Nonetheless, as shown with our *in vivo* analyses, application of the
474 EGRIN and PRIME models to new datasets offers key insights into causal mechanistic drivers of
475 adaptive strategies of the pathogen. Given that less than 10% of transcriptomic information and
476 less than 2% of ChIP-seq regulator datasets were available for CD630, as compared to EGRIN
477 models developed for other species, the model provides a formative tool to design future
478 transcriptomic and ChIP-seq studies to improve predictions for these regulons.
479
480 Leveraging additional Tn-seq and *in vivo* transcriptomic datasets, the expanded icdf838 model
481 identified a broader set of amino acids, in addition to genes and anaerobe-specific pathways,
482 needed to support colonization and growth expansion *in vivo*. Notably, predictions of *in vivo* gene
483 essentially identified folate one-carbon cycling pathways including those connected with Wood-
484 Ljundahl fixation of carbon dioxide to acetate (Gößner et al., 2008). Predictions of gene
485 essentiality also identified multiple nucleotide synthesis and salvage pathway genes that were

486 essential *in vivo* but not *in vitro,* including ones associated with xanthine transport and
487 metabolism, an abundant nucleotide in gut secretions that originates from host sources
488 (Girinathan et al., 2021). Lastly, the systems approach using the two models has identified known
489 and novel genes and TFs across disparate pathways, including sporulation, flagella biosynthesis,
490 sugar metabolism, and biosynthesis of aromatic and branched-chain amino acids, that contribute
491 to growth in an intestinal environment. Once experimentally validated, these essential genes
492 represent novel vulnerabilities that can be rationally targeted with small molecules,
493 bacteriotherapeutics, or other patient interventions. We believe that by democratizing the
494 disparate data, algorithms, and models through interactive exploration capabilities, the *C. difficile*
495 Web Portal will accelerate collaborative systems analysis of host-pathogen-commensal
496 interactions by engaging the wider scientific community.
497
498 We illustrate additional predictions from the *C. difficile* EGRIN model to enable gene- through
499 systems-level analyses of the pathogen. Though among the best described obligate anaerobes,
500 the *C. difficile* genome still contains a high number of genes of unknown function. Model
501 predictions provided new information to assign putative functions to 48 genes, including ones
502 associated with sporulation, carbohydrate transport, and other aspects of cellular metabolism.
503 Integration of the regulatory and metabolic network models with PRIME enabled unprecedented
504 insight into the essential role of TFs in mediating combinatorial control of metabolism in different
505 colonization contexts, vis-a-vis presence and absence of the protective commensal PBI. EGRIN
506 and PRIME models will be especially powerful to navigate the vast space of possible
507 combinatorial perturbations with *in silico* simulations and prioritize experiments that are most likely
508 to reveal mechanistic insights into phenotypes that emerge from the complex interplay of
509 regulatory and metabolic processes. For instance, PRIME accurately predicted enhanced
510 synergistic epistasis between the CodY and CcpA networks, a phenotype that was not attributable
511 to a single downstream metabolic gene, reaction or process. Rather, the significantly diminished

512 growth of a double knockout of CodY and CcpA (especially in mice co-colonized with PBI) appears
513 to emerge from the interplay of more than 35 TFs regulating ~80 genes catalyzing reactions
514 across more than two dozen metabolic processes. Importantly, these PRIME-derived biological
515 insights can iteratively drive hypothesis-driven characterization of mechanisms by which the
516 pathogen infects and colonizes a complex host environment. Furthermore, by uncovering
517 conditionally essential genes and pathways, PRIME can be used to understand how *C. difficile*
518 escapes therapies, and develop strategies to block these escape routes with synergistically acting
519 secondary antibiotics. The *C. difficile* Web Portal makes all of the tools and resources available
520 to the broader research community, providing a platform for collaboration and to support systems-
521 level investigations of the pathogen and its interactions with the host and commensal microbiota.

522 # STAR METHODS

523 ***C. difficile* genome annotation**

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

531

532 ***C. difficile* transcriptional compendium**

533 To generate a transcriptional compendium for *C. difficile*, required for constructing an EGRIN
534 model, a total of 151 publicly available transcriptomes of *C. difficile* 630 (identified by searching
535 for the 'Clostridioides difficile 630' term) were downloaded from the NCBI Gene Expression
536 Omnibus (GEO) repository (Barrett et al., 2012) in March 2020. Downloaded transcriptomes were
537 generated by 11 independent studies (**Table S1**). To integrate this data into a single dataset, we
538 computed the log2 fold-change of each transcriptome with respect to a control condition, as
539 performed in the generation of other transcriptional compendia (Moretto et al., 2016; Peterson et
540 al., 2014). This step was not necessary for transcriptional data collected with dual channel arrays
541 that included a normalizing control channel. The resulting transcriptional compendium contained
542 a total of 4,091 gene features and 127 conditions. The 127 conditions in the transcriptional
543 compendium were organized in 11 distinct conditional blocks (e.g., sporulation, *fur* deletion), as
544 shown in **Table S1**.

545

546 **Construction of the EGRIN model**

547 The EGRIN model for *C. difficile* was constructed in two stages. First, we used cMonkey2 (Reiss
548 et al., 2015), a biclustering algorithm, on the compiled compendium of 127 *C. difficile*
549 transcriptomes to simultaneously detect co-regulated gene modules and the conditions where co-
550 regulation occurs. cMonkey2 integrates functional annotation from the STRING database
551 (Szklarczyk et al., 2016), gene promoter sequences from the RSAT database (Nguyen et al.,
552 2018), and operon predictions from MicrobesOnline (Dehal et al., 2010) when detecting gene
553 modules. cMonkey2 was run using default parameters. Briefly, we used 2,000 iterations to
554 optimize the co-regulated gene modules, each one with 3-70 genes. In each iteration, cMonkey2
555 refined the gene modules by evaluating and modifying (if necessary) condition and gene
556 memberships. cMonkey2 biclustering approach allowed genes and conditions to be assigned to
557 a maximum of two and 204 different modules, respectively. *De novo* GRE search was performed
558 using MEME v. 4.12.0 (Bailey et al., 2015). Second, we used the Inferelator (Arrieta-Ortiz et al.,
559 2015), a network inference algorithm, to identify potential transcriptional regulators for the 406
560 gene modules generated by cMonkey2. The Inferelator uses a Bayesian Best Subset Regression
561 to estimate the magnitude and sign (activation or repression) of potential interactions between
562 TFs and gene modules. We bootstrapped the expression data (100 times) to avoid regression
563 overfitting (Arrieta-Ortiz et al., 2015). The Inferelator generates two scores for each TF-module
564 interaction, the corresponding regression coefficient (i.e., beta) and a confidence score. The
565 second score indicates the likelihood of the interaction. The final set of TF-module interactions
566 was defined as the 805 interactions with absolute beta values ≥ 0.1.
567
568 **Experimentally supported literature derived TF regulons**
569 We mined available literature to compile a list of experimentally supported targets for the 13
570 partially characterized transcriptional regulators (involved in sporulation, motility, carbon
571 metabolism, among other processes) shown on **Table S2**. The manually compiled regulons
572 represented a total of 1,349 regulatory interactions and involved 1,044 genes. Target genes

573 included in the compiled TF regulons were supported by transcriptional data, protein-DNA binding
574 data and *in silico* analysis of promoter regions (e.g., presence of known regulators DNA binding
575 motif).
576
577 **Module enrichment evaluation**
578 We used a hypergeometric test to identify modules of co-regulated genes in the EGRIN model
579 that were statistically enriched with manually compiled TF regulons (**Table S2**) or functional
580 pathways derived from curated annotation of *C. difficile* genome (Girinathan et al., 2021). Only
581 gene modules with adjusted hypergeometric test p-value ≤ 0.05 and containing five or more genes
582 from the relevant TF regulon or functional pathway were considered enriched.
583
584 **Assignment of putative functions to uncharacterized genes**
585 To predict the potential role of uncharacterized genes, gene functions were predicted based on
586 the functional enrichment of EGRIN modules (evaluated as explained above). Function
587 assignments were restricted to uncharacterized genes located in functionally enriched modules
588 in which 45% of more of the annotated members (i.e., genes with putative function) were assigned
589 to the over-represented function term. This second filter was implemented to focus on EGRIN
590 modules that were involved in a specific function. Thus, increasing the likelihood that the
591 uncharacterized genes were also involved in the same function.
592
593 **Analysis of *in vivo* data**
594 *In vivo* transcriptomic data from gnotobiotic mice mono-colonized with *C. difficile* ATCC43255 or
595 co-colonized with *P. bifermentans* or *C. sardiniense* (Girinathan et al., 2021) were analyzed as
596 previously described using the updated reference genome of ATCC43255 to extract gene
597 features for subsequent analysis with DESeq2 (Love et al., 2014).
598

599 **Metabolic model refinement and gene essentiality prediction**

600 A published genome-scale metabolic model of *C. difficile* 630 strain, icdf834 (Kashaf et al., 2017),
601 was used in this study and expanded by adding reactions required for *in vivo* survival of the
602 pathogen. The icdf834 model incorporates 1,227 metabolic reactions and 807 metabolites. The
603 metabolic reactions were mapped through gene-protein-reaction associations to 834 genes,
604 which represent 80% of 1,030 annotated metabolic genes in the CD630 genome (**Fig. 5A**). We
605 also curated pathway annotations that were incorrectly designated using default KEGG
606 annotations (Kanehisa et al., 2017). For example, most anaerobes do not utilize the tricarboxylic
607 citric acid (TCA) cycle, although some reactions, in reverse, support aspects of pyruvate,
608 succinate and oxaloacete metabolism. In the icdf834 model, we changed subsystem pathway
609 annotation of two reactions - i) acetyl-CoA:oxaloacetate C-acetyltransferase and ii) succinyl-CoA
610 synthase from TCA cycle to pyruvate metabolism and butanoate fermentation respectively
611 (**Supplementary File S2**). Similarly, we updated reactions originally assigned to gluconeogenesis
612 and the pentose phosphate pathway. We evaluated the homology of metabolic genes between
613 *C. difficile* 630 and ATCC43255 strain of *C. difficile* in order to use the icdf834 model for
614 representing the *in vivo* infection state of ATCC43255 strain. The details of 768 genes that are
615 predicted in this homology analysis is provided in **Supplementary File S2**. We then extended the
616 model by including four new genes and eight new exchange reactions that are required for the
617 growth of the pathogen in the *in vivo* micro-environment, based on KEGG annotations. We also
618 expanded the proline reductase (PR) and glycine reductase (GR) systems by adding alanine,
619 branched chain amino acids (valine, leucine and isoleucine) and aromatic amino acids (phenyl
620 alanine, tyrosine and tryptophan) as donors in the Stickland metabolism (Jackson et al., 2006),
621 increasing the total number of reactions from 1,227 to 1,273 (**Supplementary File S2**). We
622 named this expanded version of the model as “icdf838”. Then, the transcriptome of *C. difficile*
623 profiled from *in vivo* infections of specifically-colonized gnotobiotic mice (Girinathan et al., 2021)
624 was mapped onto the icdf838 model using the GIMME algorithm (Becker and Palsson, 2008).

625 This resulted in a model with 712 active reactions, with no changes in the number of genes. This
626 model represents the *in vivo* state of *C. difficile*. We applied the constraint-based method for
627 simulating the metabolic steady-state of *C. difficile* using flux-balance analysis (FBA) (Becker and
628 Palsson, 2008; Orth et al., 2010). The initial validation steps involved checking the capacity of the
629 icdf834 model to produce biomass in defined media conditions including 1) minimal medium, 2)
630 basal defined medium and 3) complex, nutrient-rich medium (compositions used according to
631 Larocque et al., 2014). Then, we tested the performance of both icdf834 and icdf838 models using
632 gene essentiality predictions by FBA. A gene was considered "essential" if its deletion reduced
633 the biomass by >65%. By this analysis, the model classified each gene as "essential" or "non-
634 essential". We compared the gene essentiality predictions from nutrient-rich media constraints
635 with the available experimental Tn-seq data (Dembek et al., 2015) and deduced the confusion
636 matrix to derive true positive rates (TPR) and false positive rates (FPR). This led to the elucidation
637 of sensitivity and specificity of the model using ROC curve analysis. We then applied the same
638 strategy and predicted the essential genes *in vivo* using FBA with the expanded and GIMME-
639 derived context-specific network, icdf838. All model simulations related to FBA were performed
640 on MATLAB_R2019a platform using the recent version of COBRA (The COnstraint-Based
641 Reconstruction and Analysis) toolbox (Heirendt et al., 2019). *In silico* gene essentiality predictions
642 were performed using the COBRA toolbox 'single-gene-deletion' function in MATLAB. The
643 illustration of essential gene regulatory network *in vivo* was deduced using BioTapestry tool
644 (Paquette et al., 2016).
645
646 **PRIME model development**
647 Three PRIME models were constructed for *C. difficile* during *in vivo* (mono- and PBI co-colonized
648 mice) and *in vitro* growth. Briefly, we first inferred a TF-gene transcriptional network using the
649 Inferelator with incorporation of TF activity, estimated using the signed (positive or negative) TF-
650 gene interaction network compiled for 13 TFs (**Table S2**), as we have previously done for other

651 species (Arrieta-Ortiz et al., 2015, 2020). To infer a global network, 288 putative transcriptional
652 regulators without known targets were also included as potential regulators. The inferred network
653 included 5,801 TF-gene interactions with absolute regression coefficients ≥ 0.1. The
654 transcriptional network was then integrated with the icdf838 metabolic model as explained in
655 (Immanuel et al., 2021). Importantly, PRIME models were made context-specific by excluding
656 metabolic reactions associated with lowly expressed genes, as explained above.
657
658 ***C. difficile* Web Portal**
659 This portal utilizes the powerful build, search, collaboration, and visualization features of the
660 Drupal content management system. With the two key features of modularity and extensibility,
661 Drupal provides a slim, powerful core that can be readily extended through custom modules and
662 easy-to-use collaborative tools to support information sharing. Based on these key features, we
663 developed this content management system into a data management, analysis, and visualization
664 framework to support *C. difficile* research.
665
666 Due to the complexity of the information provided by the genome and models, it is critical to
667 provide a user-friendly and flexible search and filtering capabilities. By taking advantage of
668 Drupal's built-in search interface and implementing Apache Solr search, we created very powerful
669 search capabilities that will query every information included in the portal database. Moreover,
670 the search interface uses "facets" to allow users to explore a collection of information by applying
671 multiple filters. This combination together with sorting enables users to start with broad searches
672 and then quickly pinpoint specific information.
673
674 In order to provide a comprehensive functional genomics resource for the *C. difficile* community,
675 genome annotations from several different sources were merged and imported into the C. difficile
676 Portal. Curated genome annotations for *Clostridium difficile* strain 630 published by Monot et al.

677 (Monot et al., 2011), were downloaded from MicroScope platform (Vallenet et al., 2017).
678 Additional functional annotations were downloaded from PATRIC (Wattam et al., 2014) and
679 Uniprot (Consortium, 2017) and merged with curated genome annotations. Overall, 4,018 genes
680 were included in the Cdiff Portal. The *C. difficile* genome included 1,030 metabolic genes, 309
681 TFs, 270 small non-coding RNAs (sRNAs) (Soutourina et al., 2013), 87 tRNAs, 32 rRNAs and 17
682 miscellaneous RNAs (miscRNAs). The genome included 1,330 genes with unknown function.
683 Furthermore, gene essentiality data from Dembek et al. (Dembek et al., 2015) was integrated with
684 gene annotations.
685

686 **ACKNOWLEDGEMENTS**

687 We thank members of the Baliga lab for critical discussions and feedback. Funding was provided
688 by National Institute of Allergy and Infectious Diseases (R01AI128215 and R01AI141953), and
689 National Science Foundation (BDI-1565166) to NSB, and the BWH Precision Medicine Institute,
690 Hatch Family Foundation, R01AI153605 and P30-DK034854 to LB. BPG was supported by T32
691 HL007627. JNW was supported by the Intramural Research Program of the National Library of
692 Medicine, National Institutes of Health. Funding was also provided by the Institut Pasteur and
693 Université de Paris to BD and JP, and the Institut Universitaire de France to OS.
694

695 **COMPETING INTEREST**

696 LB is an inventor on patents for *C. difficile* microbiota therapeutics. LB is an SAB member and
697 holds stock in ParetoBio Inc. Remaining authors declare no competing interests.
698

699 **AUTHOR CONTRIBUTIONS**

700 Conceptualization: MLAO, LB and NSB. Data curation: MLAO, SRCI, ST, BPG, JNW, ND, JP,
701 OS, BD and LB. Formal analysis: MLAO and SRCI. Investigation: MLAO, SRCI, JP, BD, LB and
702 NSB. Methodology: MLAO, SRCI, ST and WJW. Project Administration: MLAO and NSB.

Resources: JP, BD, LB and NSB. Software: MLAO, ST, WJW. Supervision: NSB. Visualization: MLAO, SRCI, ST, WJW, NSB. Writing: MLAO, SRCI, ST, LB and NSB.

DATA AND CODE AVAILABILITY

All datasets, models, software and supporting resources are available in the interactive and open access C. diff Web Portal (http://networks.systemsbiology.net/cdiff-portal/).

REFERENCES

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

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

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.

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*.

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*.

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. (2012). 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.

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

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.

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 J. *284*, 4343–4357.

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.

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

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*, 740–740.

Carter, G.P., Douce, G.R., Govind, R., Howarth, P.M., Mackin, K.E., Spencer, J., Buckley, A.M., Antunes, A., Kotsanas, D., Jenkin, G.A., et al. (2011). The anti-sigma factor TcdC modulates hypervirulence in an epidemic BI/NAP1/027 clinical isolate of *Clostridium difficile*. PLoS Pathog

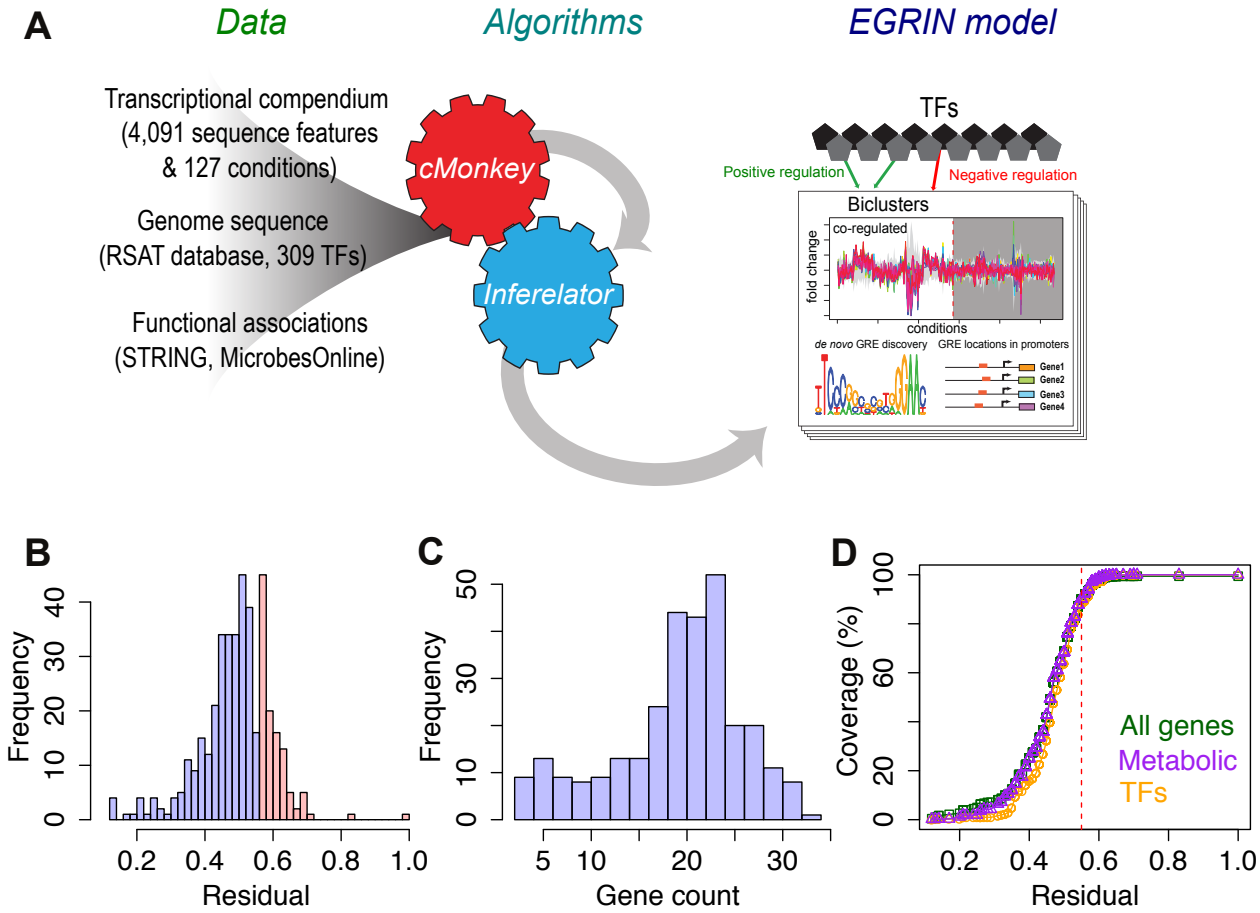
759 *7*, e1002317.
760 Cartman, S.T., Kelly, M.L., Heeg, D., Heap, J.T., and Minton, N.P. (2012). Precise manipulation
761 of the *Clostridium difficile* chromosome reveals a lack of association between the *tcdC* genotype
762 and toxin production. Appl. Environ. Microbiol. *78*, 4683–4690.
763 Consortium, T.U. (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *45*,
764 D158--D169.
765 Dehal, P.S., Joachimiak, M.P., Price, M.N., Bates, J.T., Baumohl, J.K., Chivian, D., Friedland,
766 G.D., Huang, K.H., Keller, K., Novichkov, P.S., et al. (2010). MicrobesOnline: an integrated
767 portal for comparative and functional genomics. Nucleic Acids Res. *38*, D396-400.
768 Dembek, M., Barquist, L., Boinett, C.J., Cain, A.K., Mayho, M., Lawley, T.D., Fairweather, N.F.,
769 and Fagan, R.P. (2015). High-throughput analysis of gene essentiality and sporulation in
770 *Clostridium difficile*. MBio *6*, e02383--14.
771 Dineen, S.S., Villapakkam, A.C., Nordman, J.T., and Sonenshein, A.L. (2007). Repression of
772 *Clostridium difficile* toxin gene expression by CodY. Mol. Microbiol. *66*, 206–219.
773 Dineen, S.S., McBride, S.M., and Sonenshein, A.L. (2010). Integration of metabolism and
774 virulence by *Clostridium difficile* CodY. J. Bacteriol. *192*, 5350–5362.
775 Dingle, T.C., Mulvey, G.L., and Armstrong, G.D. (2011). Mutagenic analysis of the *Clostridium*
776 *difficile* flagellar proteins, FliC and FliD, and their contribution to virulence in hamsters. Infect.
777 Immun. *79*, 4061–4067.
778 Dubois, T., Dancer-Thibonnier, M., Monot, M., Hamiot, A., Bouillaut, L., Soutourina, O., Martin-
779 Verstraete, I., and Dupuy, B. (2016). Control of *Clostridium difficile* physiopathology in response
780 to cysteine availability. Infect. Immun. *84*, 2389–2405.
781 Edwards, A.N., Tamayo, R., and McBride, S.M. (2016). A novel regulator controls *Clostridium*
782 *difficile* sporulation, motility and toxin production. Mol. Microbiol. *100*, 954–971.
783 Elena, S.F., and Lenski, R.E. (2003). Evolution experiments with microorganisms: the dynamics
784 and genetic bases of adaptation. Nat. Rev. Genet. *4*, 457–469.
785 Fimlaid, K.A., Bond, J.P., Schutz, K.C., Putnam, E.E., Leung, J.M., Lawley, T.D., and Shen, A.
786 (2013). Global analysis of the sporulation pathway of *Clostridium difficile*. PLoS Genet. *9*.
787 Galperin, M.Y., Makarova, K.S., Wolf, Y.I., and Koonin, E. V (2015). Expanded microbial
788 genome coverage and improved protein family annotation in the COG database. Nucleic Acids
789 Res. *43*, D261--D269.
790 Girinathan, B.P., DiBenedetto, N., Worley, J.N., Peltier, J., Arrieta-Ortiz, M., Immanuel, S.R.C.,
791 Lavin, Ri., Delaney, M.L., Cummins, C., Onderdonk, A.B., et al. (2021). The mechanisms of *in*
792 *vivo* commensal control of *Clostridioides difficile* virulence. BioRxiv 2020.01.04.894915.
793 Gößner, A.S., Picardal, F., Tanner, R.S., and Drake, H.L. (2008). Carbon metabolism of the
794 moderately acid-tolerant acetogen *Clostridium drakei* isolated from peat. FEMS Microbiol. Lett.
795 *287*, 236–242.
796 Govind, R., and Dupuy, B. (2012). Secretion of *Clostridium difficile* toxins A and B requires the
797 holin-like protein TcdE. PLoS Pathog *8*, e1002727.
798 Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S.N., Richelle, A., Heinken, A., Haraldsdóttir, H.S.,
799 Wachowiak, J., Keating, S.M., Vlasov, V., et al. (2019). Creation and analysis of biochemical
800 constraint-based models using the COBRA Toolbox v. 3.0. Nat. Protoc. *14*, 639–702.
801 Immanuel, S.R.C., Arrieta-Ortiz, M.L., Ruiz, R.A., Pan, M., de Lomana, A.L.G., Peterson, E.J.R.,
802 and Baliga, N.S. (2021). Quantitative prediction of conditional vulnerabilities in regulatory and
803 metabolic networks of *Mycobacterium tuberculosis*. BioRxiv.
804 Jackson, S., Calos, M., Myers, A., and Self, W.T. (2006). Analysis of proline reduction in the
805 nosocomial pathogen *Clostridium difficile*. J. Bacteriol. *188*, 8487–8495.
806 Janoir, C., Denève, C., Bouttier, S., Barbut, F., Hoys, S., Caleechum, L., Chapetón-Montes, D.,
807 Pereira, F.C., Henriques, A.O., Collignon, A., et al. (2013). Adaptive strategies and
808 pathogenesis of *Clostridium difficile* from *in vivo* transcriptomics. Infect. Immun. *81*, 3757–3769.
809 Janvilisri, T., Scaria, J., and Chang, Y.-F. (2010). Transcriptional profiling of *Clostridium difficile*

810 and Caco-2 cells during infection. J. Infect. Dis. *202*, 282–290.
811 Jenior, M.L., Leslie, J.L., Young, V.B., and Schloss, P.D. (2017). *Clostridium difficile* colonizes
812 alternative nutrient niches during infection across distinct murine gut microbiomes. Msystems *2*.
813 Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new
814 perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45*, D353--D361.
815 Karasawa, T., Ikoma, S., Yamakawa, K., and Nakamura, S. (1995). A defined growth medium
816 for *Clostridium difficile*. Microbiology *141*, 371–375.
817 Kashaf, S.S., Angione, C., and Lió, P. (2017). Making life difficult for *Clostridium difficile*:
818 augmenting the pathogen's metabolic model with transcriptomic and codon usage data for
819 better therapeutic target characterization. BMC Syst. Biol. *11*, 25.
820 Kint, N., Janoir, C., Monot, M., Hoys, S., Soutourina, O., Dupuy, B., and Martin-Verstraete, I.
821 (2017). The alternative sigma factor σ^B plays a crucial role in adaptive strategies of *Clostridium*
822 *difficile* during gut infection. Environ. Microbiol. *19*, 1933–1958.
823 Larocque, M., Chénard, T., and Najmanovich, R. (2014). A curated *C. difficile* strain 630
824 metabolic network: prediction of essential targets and inhibitors. BMC Syst. Biol. *8*, 117.
825 Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and
826 dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.
827 Mani, N., and Dupuy, B. (2001). Regulation of toxin synthesis in *Clostridium difficile* by an
828 alternative RNA polymerase sigma factor. Proc. Natl. Acad. Sci. *98*, 5844–5849.
829 Martin-Verstraete, I., Peltier, J., and Dupuy, B. (2016). The regulatory networks that control
830 *Clostridium difficile* toxin synthesis. Toxins (Basel). *8*, 153.
831 Matamouros, S., England, P., and Dupuy, B. (2007). *Clostridium difficile* toxin expression is
832 inhibited by the novel regulator TcdC. Mol. Microbiol. *64*, 1274–1288.
833 McDonald, J.A.K., Mullish, B.H., Pechlivanis, A., Liu, Z., Brignardello, J., Kao, D., Holmes, E., Li,
834 J. V, Clarke, T.B., Thursz, M.R., et al. (2018). Inhibiting growth of *Clostridioides difficile* by
835 restoring valerate, produced by the intestinal microbiota. Gastroenterology *155*, 1495–1507.
836 El Meouche, I., Peltier, J., Monot, M., Soutourina, O., Pestel-Caron, M., Dupuy, B., and Pons,
837 J.-L. (2013). Characterization of the SigD regulon of *C. difficile* and its positive control of toxin
838 production through the regulation of *tcdR*. PLoS One *8*.
839 Monegro, A.F., and Regunath, H. (2018). Hospital acquired infections. In StatPearls, (StatPearls
840 Publishing)
841 Monot, M., Boursaux-Eude, C., Thibonnier, M., Vallenet, D., Moszer, I., Medigue, C., Martin-
842 Verstraete, I., and Dupuy, B. (2011). Reannotation of the genome sequence of *Clostridium*
843 *difficile* strain 630.
844 Moretto, M., Sonego, P., Dierckxsens, N., Brilli, M., Bianco, L., Ledezma-Tejeida, D., Gama-
845 Castro, S., Galardini, M., Romualdi, C., Laukens, K., et al. (2016). COLOMBOS v3.0: leveraging
846 gene expression compendia for cross-species analyses. Nucleic Acids Res. *44*, D620-3.
847 Nawrocki, K.L., Wetzel, D., Jones, J.B., Woods, E.C., and McBride, S.M. (2018). Ethanolamine
848 is a valuable nutrient source that impacts *Clostridium difficile* pathogenesis. Environ. Microbiol.
849 *20*, 1419–1435.
850 Neumann-Schaal, M., Metzendorf, N.G., Troitzsch, D., Nuss, A.M., Hofmann, J.D., Beckstette,
851 M., Dersch, P., Otto, A., and Sievers, S. (2018). Tracking gene expression and oxidative
852 damage of O2-stressed *Clostridioides difficile* by a multi-omics approach. Anaerobe *53*, 94–107.
853 Nguyen, N.T.T., Contreras-Moreira, B., Castro-Mondragon, J.A., Santana-Garcia, W., Ossio, R.,
854 Robles-Espinoza, C.D., Bahin, M., Collombet, S., Vincens, P., Thieffry, D., et al. (2018). RSAT
855 2018: regulatory sequence analysis tools 20th anniversary. Nucleic Acids Res. *46*, W209--
856 W214.
857 Orth, J.D., Thiele, I., and Palsson, B.Ø. (2010). What is flux balance analysis? Nat. Biotechnol.
858 *28*, 245–248.
859 Paquette, S.M., Leinonen, K., and Longabaugh, W.J.R. (2016). BioTapestry now provides a
860 web application and improved drawing and layout tools. F1000Research *5*.

861 Peltier, J., Hamiot, A., Garneau, J.R., Boudry, P., Maikova, A., Hajnsdorf, E., Fortier, L.-C.,
862 Dupuy, B., and Soutourina, O. (2020). Type I toxin-antitoxin systems contribute to the
863 maintenance of mobile genetic elements in *Clostridioides difficile*. Commun. Biol. *3*, 1–13.
864 Peterson, E.J.R., Reiss, D.J., Turkarslan, S., Minch, K.J., Rustad, T., Plaisier, C.L.,
865 Longabaugh, W.J.R., Sherman, D.R., and Baliga, N.S. (2014). A high-resolution network model
866 for global gene regulation in *Mycobacterium tuberculosis*. Nucleic Acids Res. *42*, 11291–11303.
867 Reiss, D.J., Plaisier, C.L., Wu, W.-J., and Baliga, N.S. (2015). cMonkey2: Automated,
868 systematic, integrated detection of co-regulated gene modules for any organism. Nucleic Acids
869 Res. *43*, e87.
870 Riedel, T., Bunk, B., Thürmer, A., Spröer, C., Brzuszkiewicz, E., Abt, B., Gronow, S., Liesegang,
871 H., Daniel, R., and Overmann, J. (2015). Genome resequencing of the virulent and multidrug-
872 resistant reference strain *Clostridium difficile* 630. Genome Announc. *3*, e00276--15.
873 Saujet, L., Monot, M., Dupuy, B., Soutourina, O., and Martin-Verstraete, I. (2011). The key
874 sigma factor of transition phase, SigH, controls sporulation, metabolism, and virulence factor
875 expression in *Clostridium difficile*. J. Bacteriol. *193*, 3186–3196.
876 Saujet, L., Pereira, F.C., Serrano, M., Soutourina, O., Monot, M., Shelyakin, P. V, Gelfand, M.S.,
877 Dupuy, B., Henriques, A.O., and Martin-Verstraete, I. (2013). Genome-wide analysis of cell
878 type-specific gene transcription during spore formation in *Clostridium difficile*. PLoS Genet *9*,
879 e1003756.
880 Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–
881 2069.
882 Soutourina, O., Dubois, T., Monot, M., Shelyakin, P. V, Saujet, L., Boudry, P., Gelfand, M.S.,
883 Dupuy, B., and Martin-Verstraete, I. (2020). Genome-Wide Transcription Start Site Mapping and
884 Promoter Assignments to a Sigma Factor in the Human Enteropathogen *Clostridioides difficile*.
885 Front. Microbiol. *11*, 1939.
886 Soutourina, O.A., Monot, M., Boudry, P., Saujet, L., Pichon, C., Sismeiro, O., Semenova, E.,
887 Severinov, K., Le Bouguenec, C., Coppée, J.-Y., et al. (2013). Genome-wide identification of
888 regulatory RNAs in the human pathogen *Clostridium difficile*. PLoS Genet *9*, e1003493.
889 Steglich, M., Hofmann, J.D., Helmecke, J., Sikorski, J., Spröer, C., Riedel, T., Bunk, B.,
890 Overmann, J., Neumann-Schaal, M., and Nübel, U. (2018). Convergent loss of ABC transporter
891 genes from *Clostridioides difficile* genomes is associated with impaired tyrosine uptake and p-
892 cresol production. Front. Microbiol. *9*, 901.
893 Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A.,
894 Doncheva, N.T., Roth, A., Bork, P., et al. (2016). The STRING database in 2017: quality-
895 controlled protein--protein association networks, made broadly accessible. Nucleic Acids Res.
896 gkw937.
897 Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L.,
898 Lomsadze, A., Pruitt, K.D., Borodovsky, M., and Ostell, J. (2016). NCBI prokaryotic genome
899 annotation pipeline. Nucleic Acids Res. *44*, 6614–6624.
900 Theriot, C.M., Koenigsknecht, M.J., Carlson Jr, P.E., Hatton, G.E., Nelson, A.M., Li, B.,
901 Huffnagle, G.B., Li, J.Z., and Young, V.B. (2014). Antibiotic-induced shifts in the mouse gut
902 microbiome and metabolome increase susceptibility to *Clostridium difficile* infection. Nat.
903 Commun. *5*, 3114.
904 Underwood, S., Guan, S., Vijayasubhash, V., Baines, S.D., Graham, L., Lewis, R.J., Wilcox,
905 M.H., and Stephenson, K. (2009). Characterization of the sporulation initiation pathway of
906 *Clostridium difficile* and its role in toxin production. J. Bacteriol. *191*, 7296–7305.
907 Vallenet, D., Calteau, A., Cruveiller, S., Gachet, M., Lajus, A., Josso, A., Mercier, J., Renaux, A.,
908 Rollin, J., Rouy, Z., et al. (2017). MicroScope in 2017: an expanding and evolving integrated
909 resource for community expertise of microbial genomes. Nucleic Acids Res. *45*, D517--D528.
910 Vemuri, R.C., Gundamaraju, R., Shinde, T., and Eri, R. (2017). Therapeutic interventions for gut
911 dysbiosis and related disorders in the elderly: antibiotics, probiotics or faecal microbiota

912 transplantation? Benef. Microbes *8*, 179–192.
913 Walter, B.M., Rupnik, M., Hodnik, V., Anderluh, G., Dupuy, B., Paulič, N., Žgur-Bertok, D., and
914 Butala, M. (2014). The LexA regulated genes of the *Clostridium difficile*. BMC Microbiol. *14*, 88.
915 Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J.,
916 Gough, R., Hix, D., Kenyon, R., et al. (2014). PATRIC, the bacterial bioinformatics database and
917 analysis resource. Nucleic Acids Res. *42*, D581--D591.
918 Wolfe, C.J., Kohane, I.S., and Butte, A.J. (2005). Systematic survey reveals general applicability
919 of" guilt-by-association" within gene coexpression networks. BMC Bioinformatics *6*, 227.
920 Woods, E.C., Nawrocki, K.L., Suárez, J.M., and McBride, S.M. (2016). The *Clostridium difficile*
921 Dlt pathway is controlled by the extracytoplasmic function sigma factor σ^V in response to
922 lysozyme. Infect. Immun. *84*, 1902–1916.
923
924

925 **FIGURES**

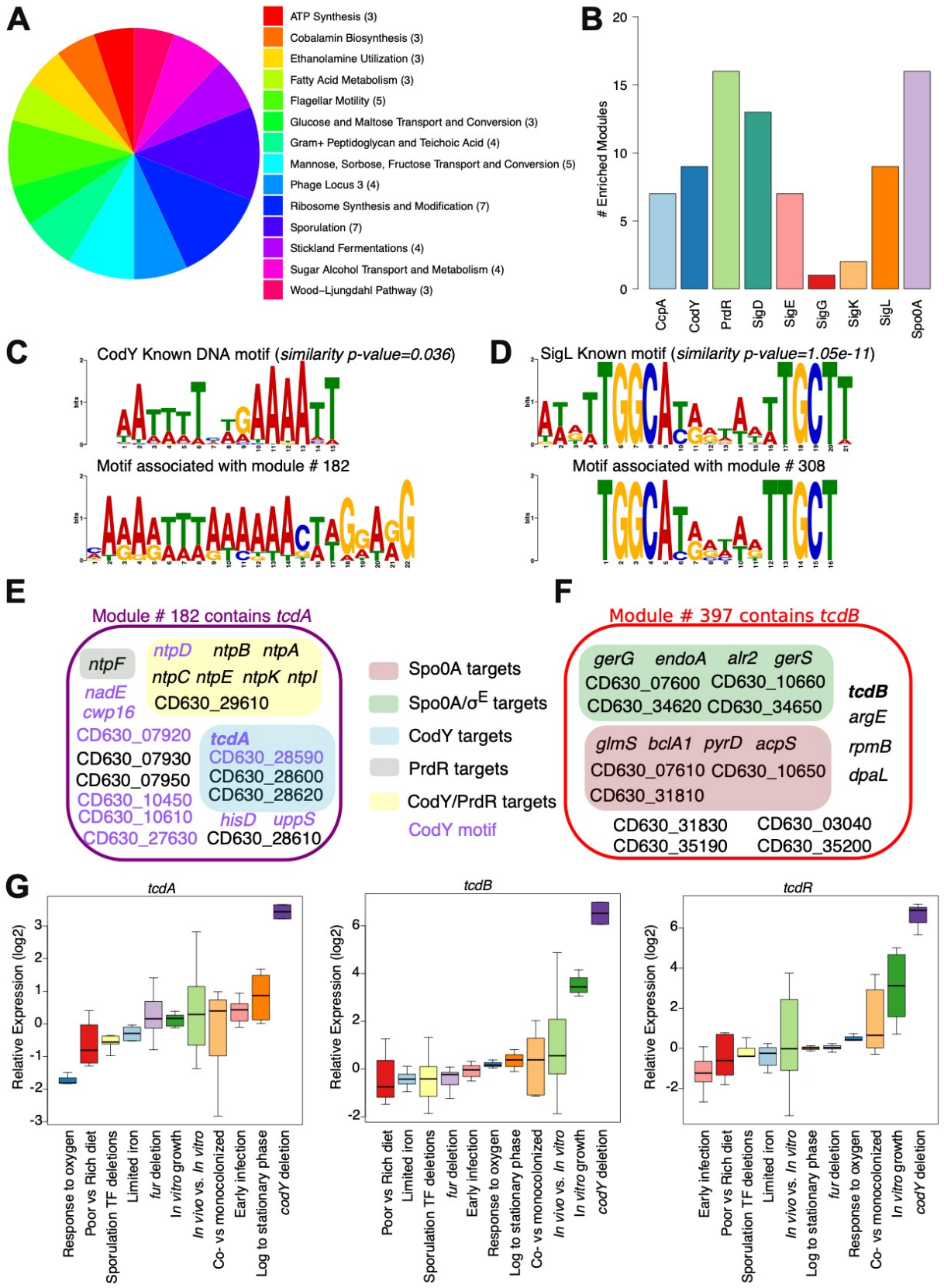


926

927 **Figure 1. Inference pipeline and general properties of the resulting Environment and Gene**
928 **Regulatory Influence Network (EGRIN) model of *C. difficile*.** (A) Framework used to build the
929 EGRIN model. (B) Distribution of residual values for the 406 detected co-regulated gene modules.
930 297 gene modules had residual values equal or lower than 0.55 (shown in purple) and were
931 labelled as high quality. (C) Distribution of gene count for the high quality gene modules. (D)
932 Coverage of all genes (4,018), the subset of metabolic genes (1,030) and TFs (309) by EGRIN
933 modules for different residual thresholds. The red dashed line indicates the 0.55 residual cutoff.

934

935


A
ATP Synthesis (3)
Cobalamin Biosynthesis (3)
Ethanolamine Utilization (3)
Fatty Acid Metabolism (3)
Flagellar Motility (5)
Glucose and Maltose Transport and Conversion (3)
Gram+ Peptidoglycan and Teichoic Acid (4)
Mannose, Sorbose, Fructose Transport and Conversion (5)
Phage Locus 3 (4)
Ribosome Synthesis and Modification (7)
Sporulation (7)
Stickland Fermentations (4)
Sugar Alcohol Transport and Metabolism (4)
Wood–Ljungdahl Pathway (3)
B
Enriched Modules
CcpA
CodY
PrdR
SigD
SigE
SigG
SigK
SigL
Spo0A
C
CodY Known DNA motif (similarity p-value=0.036)
Motif associated with module # 182
D
SigL Known motif (similarity p-value=1.05e-11)
Motif associated with module # 308
E
Module # 182 contains tcdA
ntpF
nadE
cwp16
CD630_07920
CD630_07930
CD630_07950
CD630_10450
CD630_10610
CD630_27630
ntpD ntpB ntpA
ntpC ntpE ntpK ntpI
CD630_29610
tcdA
CD630_28590
CD630_28600
CD630_28620
hisD uppS
CD630_28610
Spo0A targets
Spo0A/σE targets
CodY targets
PrdR targets
CodY/PrdR targets
CodY motif
F
Module # 397 contains tcdB
gerG endoA alr2 gerS
CD630_07600 CD630_10660
CD630_34620 CD630_34650
glmS bclA1 pyrD acpS
CD630_07610 CD630_10650
CD630_31810
tcdB
argE
rpmB
dpaL
CD630_31830 CD630_03040
CD630_35190 CD630_35200
G
tcdA
tcdB
tcdR
Relative Expression (log2)
Response to oxygen
Poor vs Rich diet
Sporulation TF deletions
Limited iron
fur deletion
In vitro growth
In vivo vs. In vitro
Co- vs monocolonized
Early infection
Log to stationary phase
codY deletion

936

937 **Figure 2. The Environment and Gene Regulatory Influence Network model of *C. difficile***
938 **recapitulates known biology of the pathogen.** (A) Co-regulated gene modules are enriched
939 with functional terms derived from expert curated annotation of the *C. difficile* genome (Girinathan
940 et al., 2021). The pie chart shows terms over-represented in three or more modules. Number of
941 modules associated with each functional term is shown in parenthesis. (B) Enriched gene
942 modules among nine (out of 13) manually-defined and experimentally supported TF regulons
943 (compiled from publicly available data in Table S2). (C) EGRIN identified the known DNA binding
944 motif of CodY (Dineen et al., 2010). (D) EGRIN also identified the known DNA binding motif of
945 SigL (Soutourina et al., 2020). Motif comparisons were performed using Tomtom (Bailey et al.,
946 2015). (E) The EGRIN model recapitulated the previously reported influence of CodY on *tcdA*
947 expression. The module #182 contains *tcdA*, it is enriched with members of the CodY regulon
948 and contains a GRE (shown in panel C) similar to the experimentally determined CodY motif. (F)
949 The EGRIN model also captured the interaction between toxin expression and sporulation via
950 module #397 that contains *tcdB* and is enriched with genes regulated by sporulation-related
951 transcriptional regulators. **(**G) Expression profiles of *tcdAB* and *tcdR* (positive regulator of the
952 pathogenicity loci). Highest expression of the toxin genes and their activator was observed in the
953 *codY* single deletion condition (dark purple box).
954

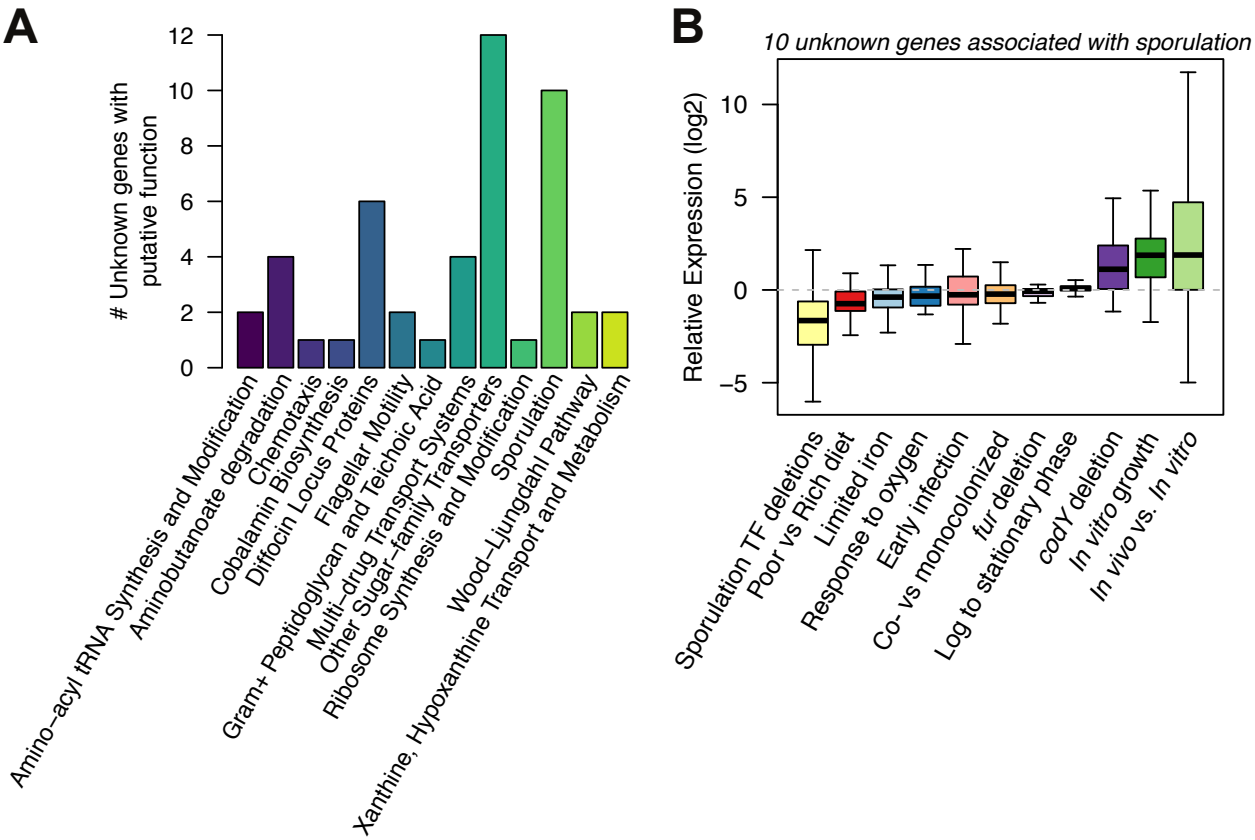
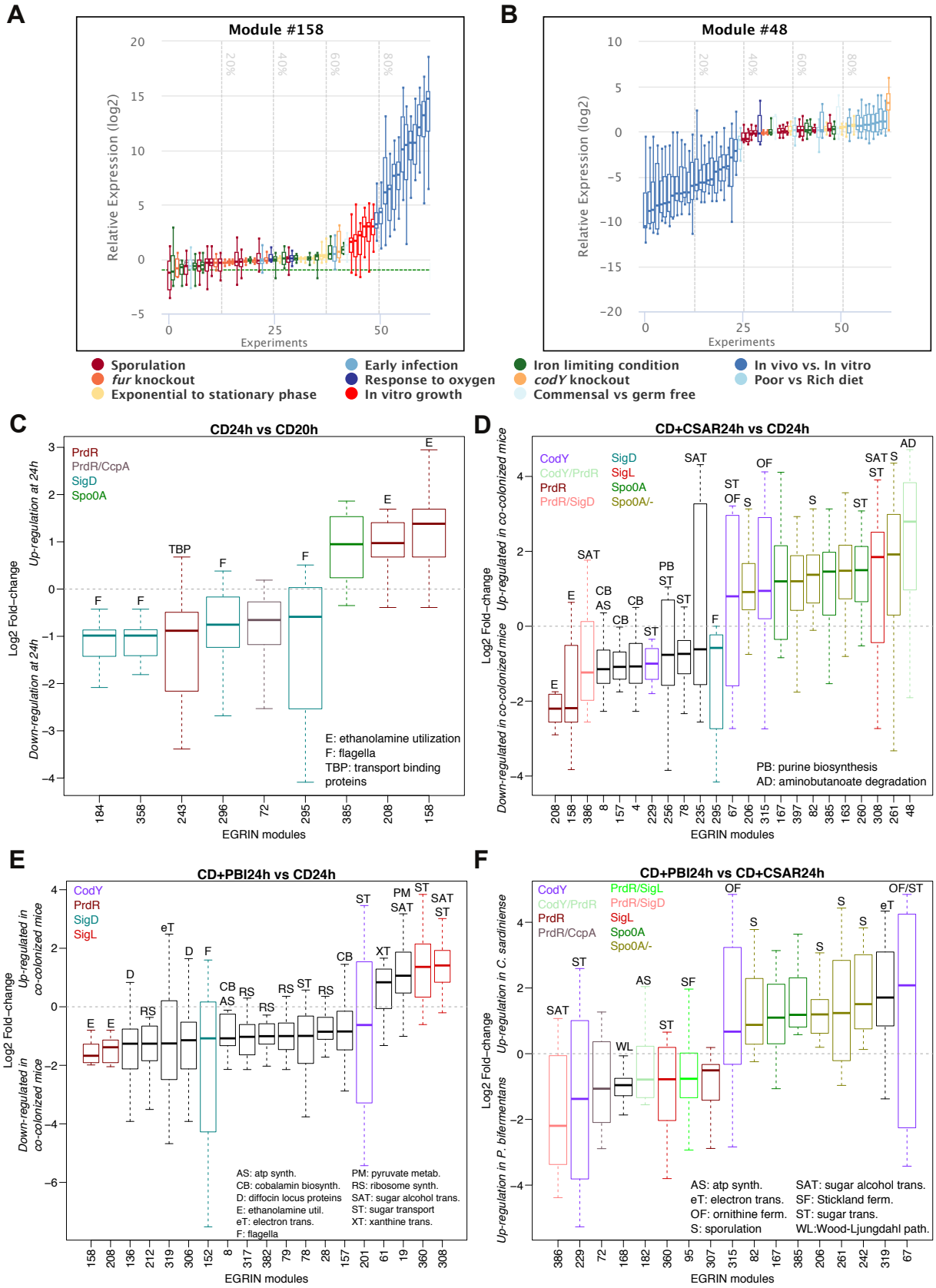


Figure 3. The EGRIN model offers insights on potential functions of uncharacterized genes of *C. difficile*. Hypotheses regarding the functions of 48 uncharacterized genes were generated based on their membership in high quality co-regulated gene modules significantly enriched with specific functional 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 the observed strongest and significant downregulation in single deletion strains of transcriptional regulators of sporulation (*spo0A*, *sigEFGK*, *spoIIID*).


A
Module #158
Relative Expression (log2)
Experiments
B
Module #48
Relative Expression (log2)
Experiments
Sporulation
fur knockout
Exponential to stationary phase
Early infection
Response to oxygen
In vitro growth
Iron limiting condition
codY knockout
Commensal vs germ free
In vivo vs. In vitro
Poor vs Rich diet
C
CD24h vs CD20h
PrdR
PrdR/CcpA
SigD
Spo0A
Log2 Fold-change
Up-regulation at 24h
Down-regulation at 24h
E: ethanolamine utilization
F: flagella
TBP: transport binding proteins
EGRIN modules
D
CD+CSAR24h vs CD24h
CodY
CodY/PrdR
PrdR
PrdR/SigD
SigD
SigL
Spo0A
Spo0A/-
Log2 Fold-change
Up-regulated in co-colonized mice
Down-regulated in co-colonized mice
PB: purine biosynthesis
AD: aminobutanoate degradation
EGRIN modules
E
CD+PBI24h vs CD24h
CodY
PrdR
SigD
SigL
Log2 Fold-change
Up-regulated in co-colonized mice
Down-regulated in co-colonized mice
AS: atp synth.
CB: cobalamin biosynth.
D: diffocin locus proteins
E: ethanolamine util.
eT: electron trans.
F: flagella
PM: pyruvate metab.
RS: ribosome synth.
SAT: sugar alcohol trans.
ST: sugar transport
XT: xanthine trans.
EGRIN modules
F
CD+PBI24h vs CD+CSAR24h
CodY
CodY/PrdR
PrdR
PrdR/CcpA
PrdR/SigL
PrdR/SigD
SigL
Spo0A
Spo0A/-
Log2 Fold-change
Up-regulation in C. sardiniense
Up-regulation in P. bifermentans
AS: atp synth.
eT: electron trans.
OF: ornithine ferm.
S: sporulation
SAT: sugar alcohol trans.
SF: Stickland ferm.
ST: sugar trans.
WL:Wood-Ljungdahl path.
EGRIN modules

965

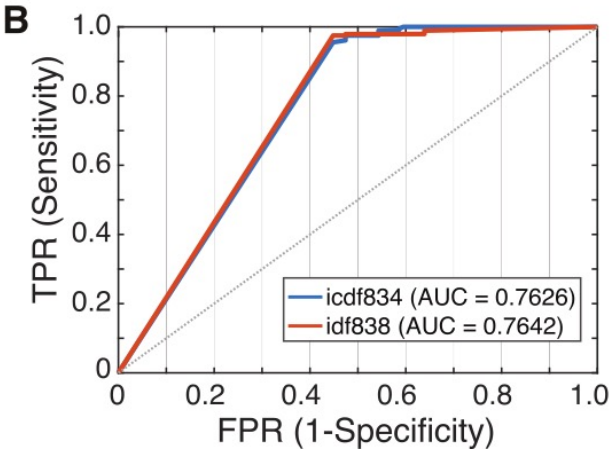
966 **Figure 4. The EGRIN model identifies TFs driving the *in vivo* response of *C. difficile* when**
967 **interacting with gut commensals *P. bifermentans* (PBI) and *C. sardiniense* (CSAR).** (A)
968 Expression profile of module #158 across the transcriptional compendium used to build the
969 EGRIN model. Each box represents the log2 fold-change of all members of the module in a single
970 transcriptome. Transcriptomes are color coded according to their membership in the 11 condition
971 blocks included in the compendium. Transcriptomes were ranked based on their median log2
972 fold-change. Only one condition block ('In vivo vs. in vitro') was found in the 20% of highest
973 (indicated with the 80% dashed line) fold-change. (B) Expression profile of module #48. The 'Early
974 infection' condition block is statistically over-represented in the 20% of highest transcriptomes
975 median log2 fold-change. (C) EGRIN modules enriched with genes differentially expressed
976 (absolute log2 fold-change > 1 and adjusted p-value < 0.05) in *C. difficile* mono-colonized mice
977 at 24 vs 20 hours of infection. X-axis shows module IDs. Modules were annotated according to
978 their functional enrichment and overlap with manually curated TF regulons (Table S2). (D)
979 Enriched EGRIN modules in *C. sardiniense*+*C. difficile* co-colonized mice vs *C. difficile* mono-
980 colonized mice at 24 hours of infection. Due to space constraint, only abbreviations of functional
981 terms not shown in other panels are displayed. (E) Enriched EGRIN modules in *P.*
982 *bifermentans*+*C. difficile* co-colonized mice vs *C. difficile* mono-colonized mice at 24 hours of
983 infection. (F) Enriched EGRIN modules in *P. bifermentans*+*C. difficile* co-colonized mice vs *C.*
984 *sardiniense*+*C. difficile* co-colonized mice at 24 hours of infection. For all comparisons, only
985 modules with absolute median fold-changes > 0.5, and enriched with TF regulons or functional
986 categories are displayed.
987
988
989

A

In vitro constraints-based metabolic model of *C. difficile*	*icdf834*	*icdf838*
Genes	834	838
Reactions	1227	1273
Metabolites	807	812

C

In vivo constraints-based metabolic model of *C. difficile*	
Genes	838
Reactions	712
Metabolites	607



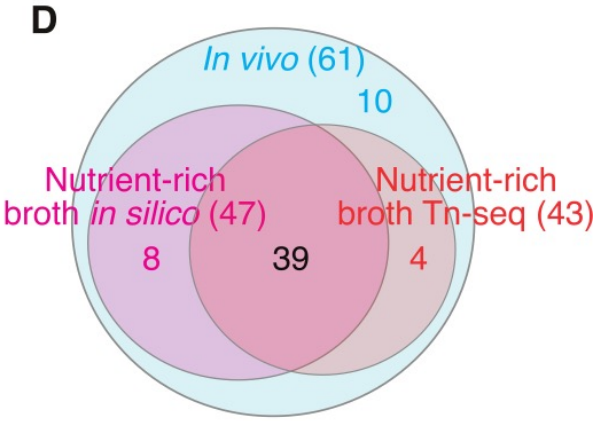
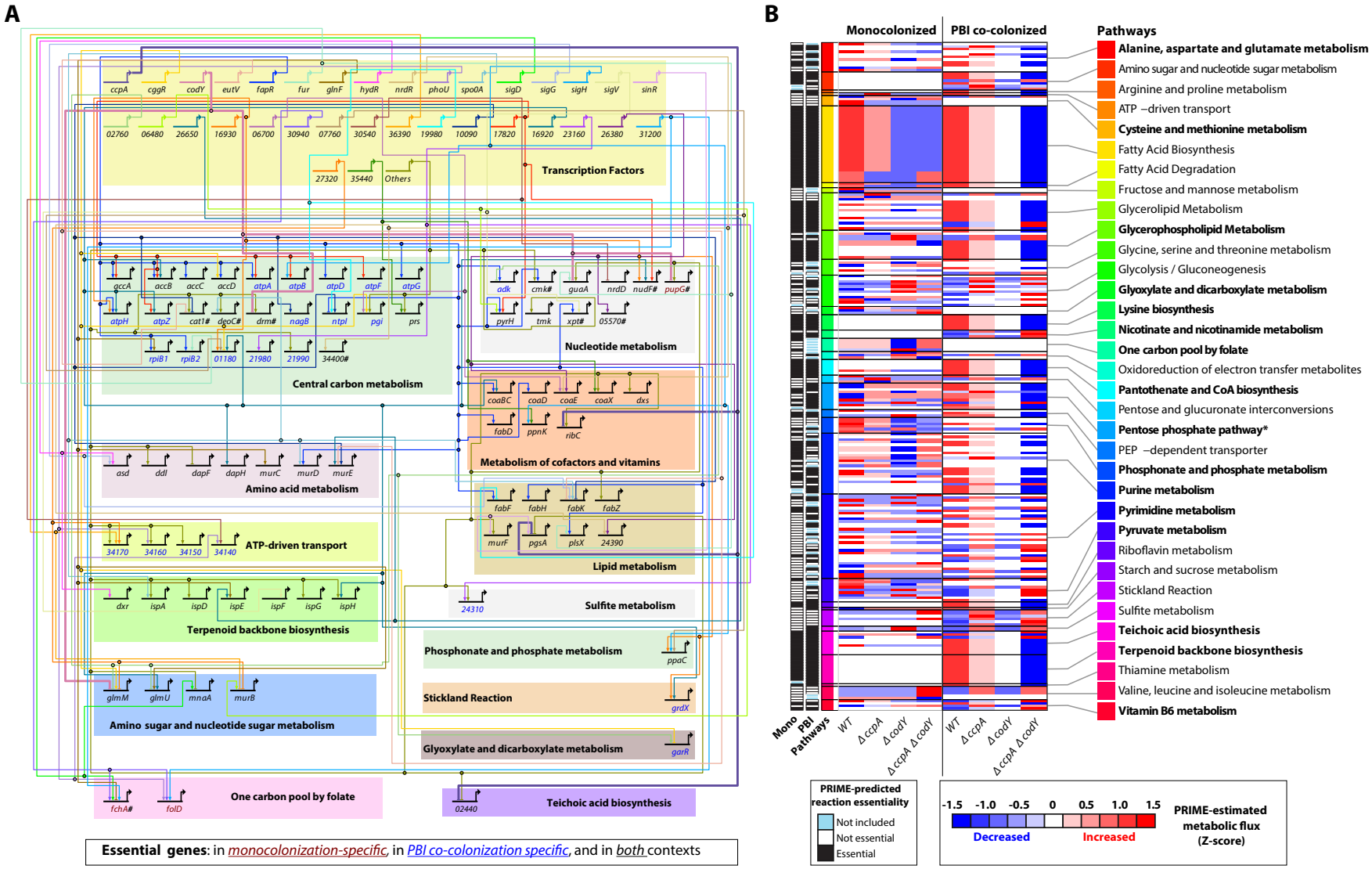


Figure 5. Metabolic model predictions. (A) Details of the *in vitro* metabolic models (icdf834 and icdf838) of *C. difficile* 630 (Kashaf et al., 2017). The number of genes, reactions and metabolites in the icdf838 model (generated in this study) after adding the required *in vivo* exchanges, transports and new reactions are shown. (B) ROC curves showing the accuracy of icdf834- and icdf838-predicted gene essentiality in nutrient rich medium evaluated against a Tn-seq functional screen (Dembek et al., 2015). (C) Details of the *in vivo* model derived using the GIMME algorithm (Becker and Palsson, 2008) on the icdf838 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* vs *in vivo* (i.e., mono-colonized condition).


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
gccA 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
dxr ispA ispD ispE ispF ispG ispH
Terpenoid backbone biosynthesis
24310
Sulfite metabolism
Phosphonate and phosphate metabolism
ppaC
glmM glmU mnaA murB
Amino sugar and nucleotide sugar metabolism
Stickland Reaction
grdX
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 phosphate pathway*
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
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)

1002 **Figure 6. PRIME-estimated metabolic fluxes and gene essentiality in mono- and PBI co-**
1003 **colonized conditions.** (A) BioTapestry visualization of *in vivo* gene regulatory network for *C.*
1004 *difficile* 630: 79 essential genes with predicted transcriptional regulators are shown.
1005 Transcriptional interactions were predicted using the Inferelator. The genes and regulators shown
1006 as five digit numbers represent the nomenclature preceded by 'CD630_'. All regulators with less
1007 than three essential targets were combined in the 'Others' meta-regulator. The '#' symbol
1008 indicates the 10 genes that are essential in the mono-colonized condition but not essential for *in*
1009 *vitro* growth. (B) PRIME-estimated metabolic fluxes for each reaction in the mono- and co-
1010 colonized conditions. Z-score transformation was independently applied to each reaction and
1011 condition. The '*' symbol indicates that there is some contention about the existence of this
1012 pathway in *C. difficile* and it may need to be updated in future model revisions.
1013

1014 # TABLES

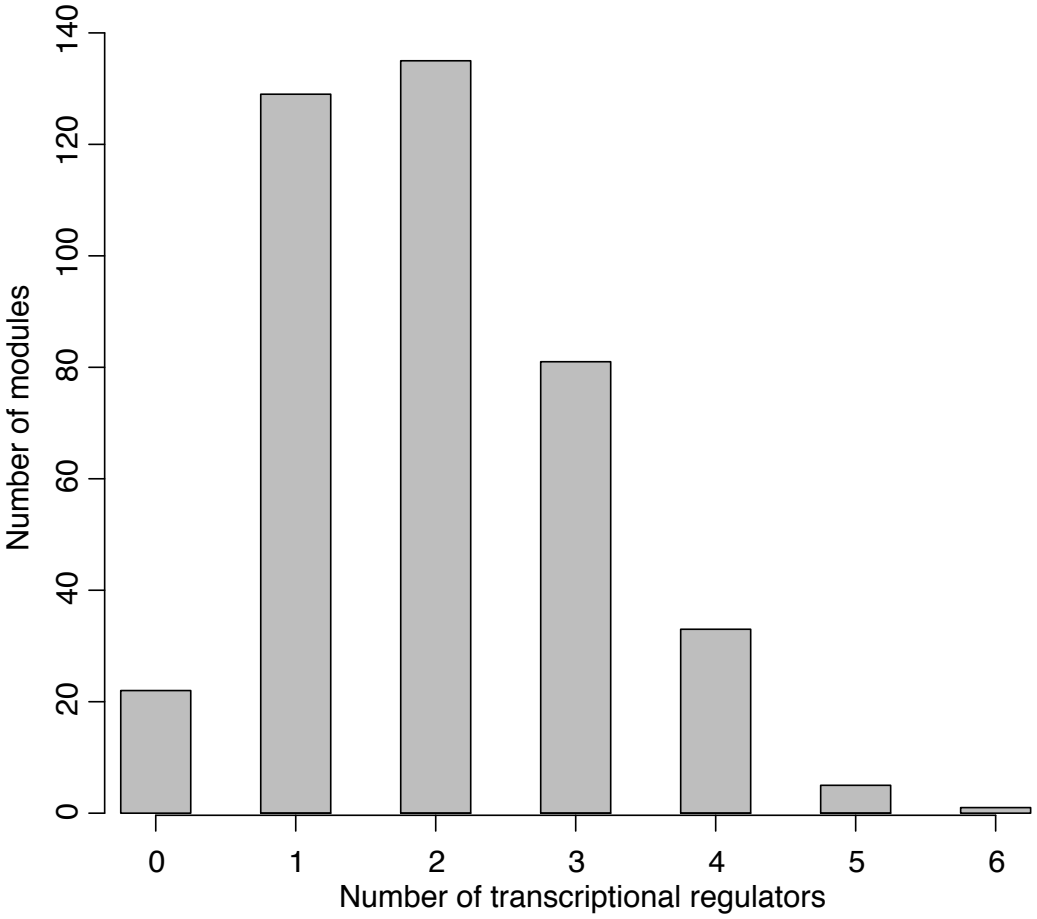
1015 **Table 1. Experimental validation of PRIME-predicted synergistic epistasis between CcpA**
1016 **and CodY.**

TF KO Strain	Mono-colonized				PBI co-colonized			
	PRIME prediction		Experiment		PRIME prediction		Experiment	
	Relative growth (Log10)[a]	Essentiality[b]	Relative growth (Log10)[a,c] 24 h	Essentiality[b]	Relative growth (Log10)[a]	Essentiality[b]	Relative growth (Log10)[a,c] 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

1017 [a]Relative growth with respect to wild type genotype
1018 [b]Gene deletions that reduce *C. difficile* growth by more than 65% (i.e., log10(relative growth) < -
1019 0.456) were labelled as essential
1020 [c]T-test P-values ≥ 0.05 are indicated with the '#' symbol and were considered not significant.
1021 The '*', '**' and '***' symbols indicate P-values < 0.05, < 0.01 and < 0.001, respectively.

1022

1023 **SUPPLEMENTARY FIGURES**

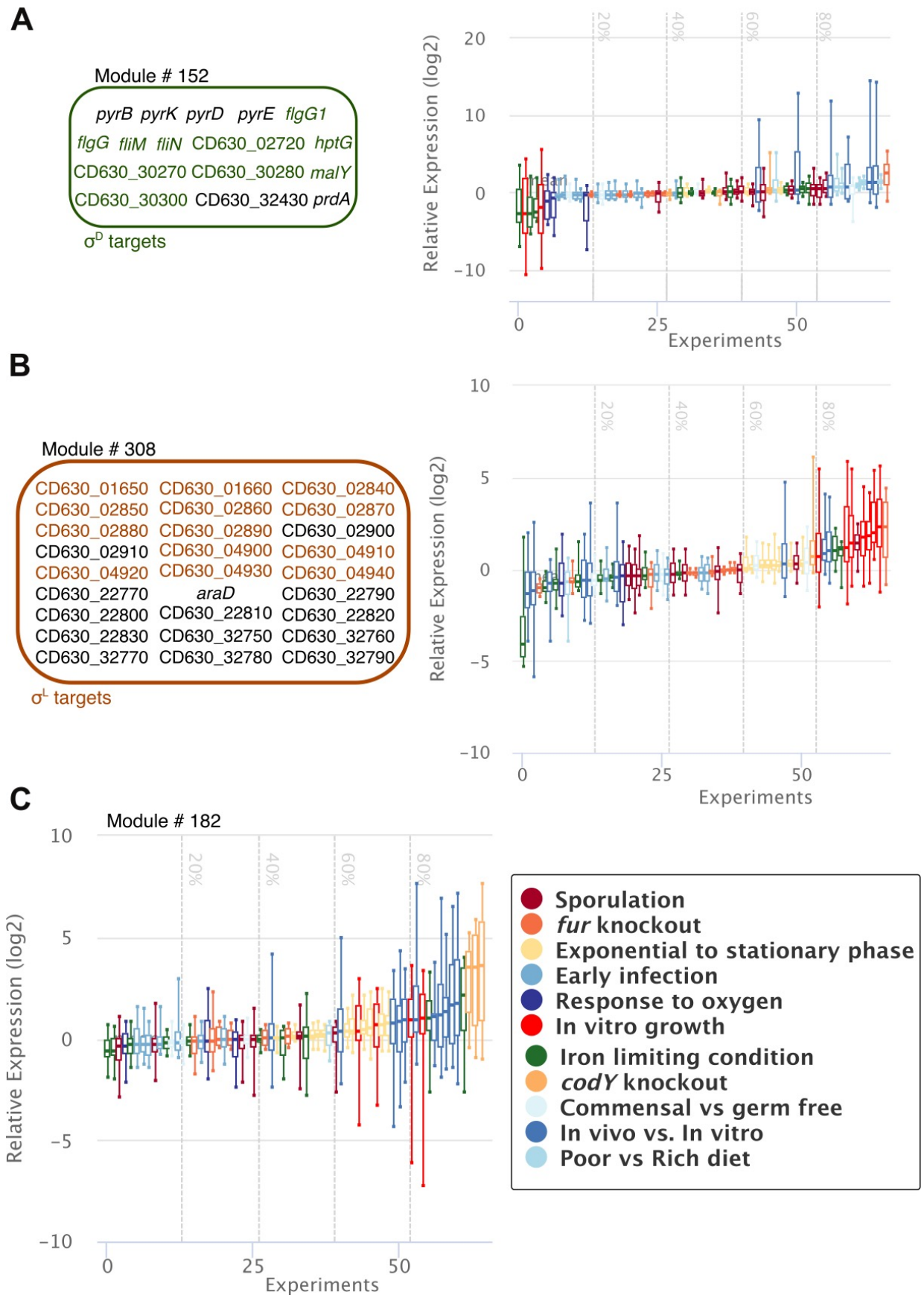


1024

1025 **Figure S1. Number of Inferelator-predicted transcriptional regulators of gene modules in**

1026 **the EGRIN model.**

1027


A
Module # 152
pyrB pyrK pyrD pyrE flgG1
flgG fliM fliN CD630_02720 hptG
CD630_30270 CD630_30280 malY
CD630_30300 CD630_32430 prdA
σD targets
Relative Expression (log2)
Experiments
20%
40%
60%
80%
B
Module # 308
CD630_01650 CD630_01660 CD630_02840
CD630_02850 CD630_02860 CD630_02870
CD630_02880 CD630_02890 CD630_02900
CD630_02910 CD630_04900 CD630_04910
CD630_04920 CD630_04930 CD630_04940
CD630_22770 araD CD630_22790
CD630_22800 CD630_22810 CD630_22820
CD630_22830 CD630_32750 CD630_32760
CD630_32770 CD630_32780 CD630_32790
σL targets
C
Module # 182
Sporulation
fur knockout
Exponential to stationary phase
Early infection
Response to oxygen
In vitro growth
Iron limiting condition
codY knockout
Commensal vs germ free
In vivo vs. In vitro
Poor vs Rich diet

1028

1029 **Figure S2. General properties of selected EGRIN modules.** (A) Members of module # 152 are
1030 shown in the left panel. Transcriptional profile of module # 152 is shown in the right panel. Each
1031 box represents the log2 fold-change of all members of the module in a single transcriptome.
1032 Transcriptomes are color coded according to their membership in the 11 condition blocks included
1033 in the compendium (shown in the figure key). Transcriptomes were ranked based on their median
1034 log2 fold-change. (B) Members of module # 308 are shown in the left panel. Transcriptional profile
1035 of module # 308 is shown in the right panel. (C) Transcriptional profile of module # 182.

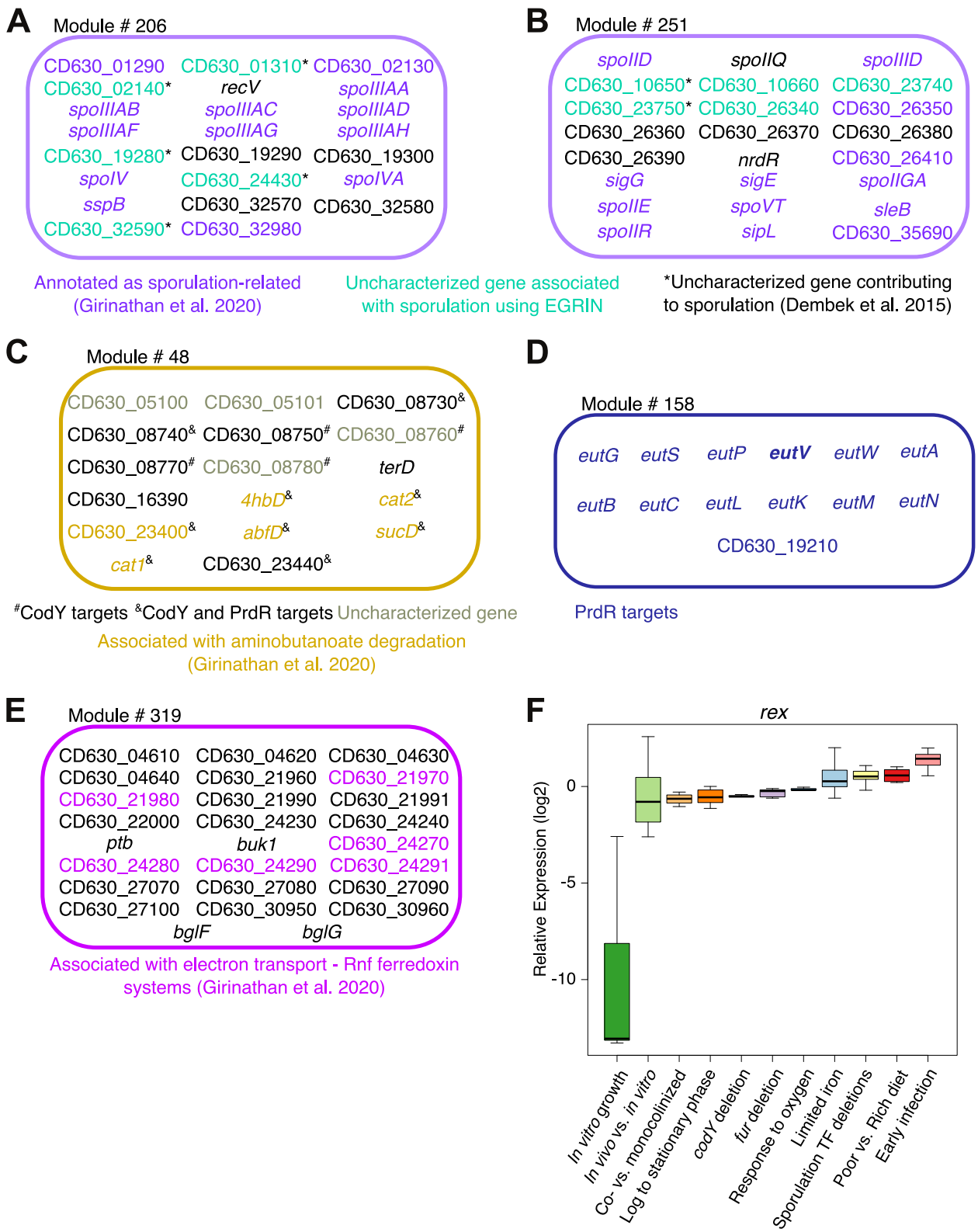
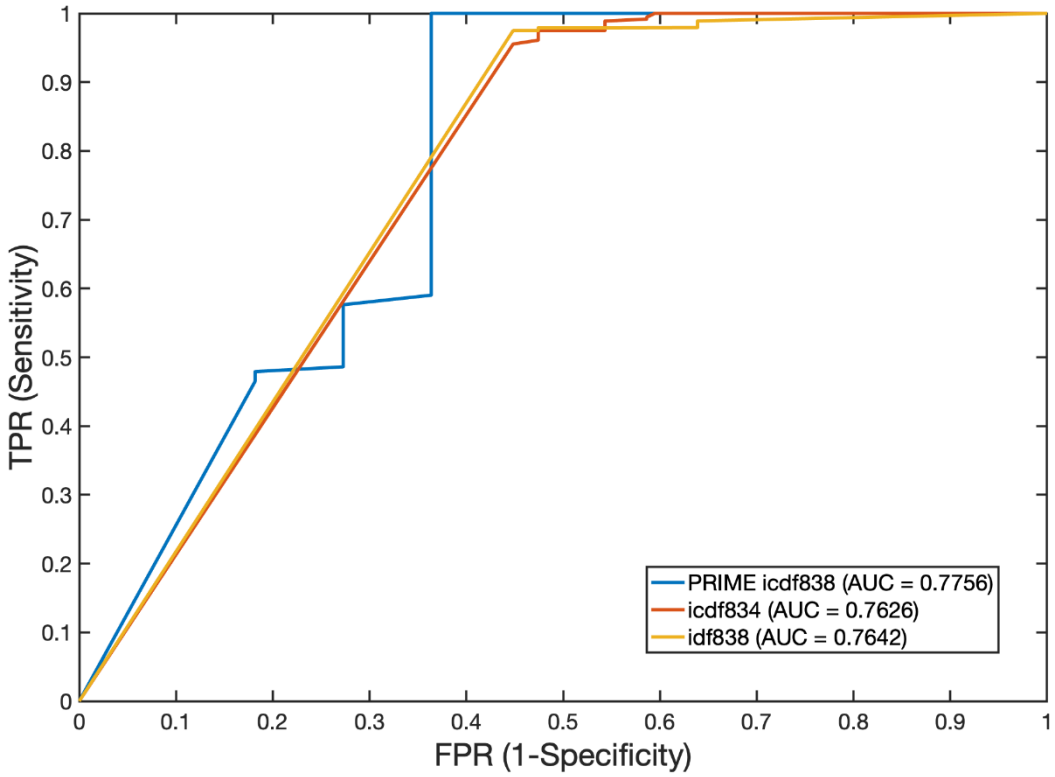


Figure S3. Modules associated with uncharacterized genes with putatively assigned function and *in vivo* adaptation. (A) Members of module # 206 are shown. (B) Members of

1039 module # 251 are shown. (C) Members of module # 48 are shown. (D) Members of module #158
1040 are shown. (E) Members of module #319 are shown. (F) Transcriptional profile of *rex* in the
1041 transcriptional compendium compiled to build the EGRIN model.

1042



1043

1044 **Figure S4. PRIME accurately predicts TF and gene essentiality.** ROC curves showing the
1045 accuracy of icdf834-, icdf838 and PRIME icdf838-predicted gene essentiality in nutrient rich
1046 medium evaluated against a Tn-seq functional screen (Dembek et al., 2015). Area Under the
1047 Curves (AUC) values for all models are shown.

A
C.diff Portal
Home
Genes
Gene Regulatory Network
Metabolic Network
About
Resources
Clostridioides difficile causes >500,000 infections, >30,000 deaths, and > $5 billion/year in US healthcare costs, and rates continue to rise. Antimicrobial therapy that ablates the commensal microbiota commonly triggers infection, allowing the pathogen to proliferate and release toxins that damage host mucosal surfaces. Resistance to commonly prescribed antibiotics occurs in >20% of patient isolates, a problem that confounds treatment and increases risks for recurrent infections. In this research program, we are building predictive and mechanistic models to define mechanisms by which C. difficile responds to antibiotics, host and microbiota-origin factors, and to develop therapeutic interventions to prevent colonization, infection and recurrence of C. difficile.
Genes
Regulatory Modules
Metabolic Reactions
Search
Enter your search term e.g "CD630_11990", "spoIIIAH", "CD630_20320", "ArgB", "Acetylglutamate kinase", "GO:0003755" etc.
Put quotes around phrases to match all the words: "fatty acid". You can require or exclude terms using + and –
Module Enrichment Analysis
You can use this tool to quickly check enrichment of your gene set in any of the regulatory network modules.
Count
Genes (4,018)
Metabolic Reactions (1,227)
Regulatory Modules (406)
miscRNA (17)
plasmid (21)
sRNA (270)
Metabolic Genes (1,030)
TF (309)
tRNA (87)
rRNA (32)
Unknown function (1,330)
Others (1,230)
B
Module 182
Summary
Residual
Gene Count
Condition Count
0.56
24
65
Bicluster expression profile
Relative Expression (log2)
Highcharts.com
Module Enrichment profile
Enrichment of Condition Blocks in Quintiles (–log10 P-value)
Highcharts.com
Sporulation
fur knockout
Exponential to stationary phase
Early infection
Response to oxygen
In vitro growth
Iron limiting condition
codY knockout
Commensal vs germ free
In vivo vs. In vitro
Poor vs Rich diet
Motif 1
Motif 1 e-value
CAAAATTTAAAAAACTAGGAGG
0.000000000023
Motif 2
Motif 2 e-value
GAAAGGTGATT
0.06
This module is predicted to be regulated by the following regulators by using the inferelator
Beta
0.1
Apply
Default filter:>=0.1 (Absolute Beta)
TF
Module
Confidence Score
Beta
CD630_19980
Transcriptional regulator, TetR family
182
0.37
-0.1690
CD630_11770
fapR
Transcriptional regulator, DeoR family (Fatty acid and phospholipid biosynthesis regulator)
182
0.48
0.2902

1048

1049 **Figure S5. The Cdiff Web Portal.** (A) Home page of the Cdiff Portal. Users can explore the gene
1050 regulatory network model and the metabolic model for *C. difficile*. In addition, all files are
1051 accessible in the resource tab. The search bar facilitates website exploration. (B) Module #182,
1052 associated with CodY and shown in Fig 2D is used as an example. Each module page includes
1053 general statistics of the module (residual score, gene count), displays the module expression
1054 profile in the compiled transcriptional compendium, the enrichment of each quintile of the module
1055 expression profile with the different condition blocks (e.g., early infection, etc.), and the detected
1056 GREs. A module page also offers information about the potential transcriptional regulators of the
1057 module. Putative regulators are defined based on over-representation of manually compiled TF
1058 regulons (assessed using hypergeometric test) and based on the Inferelator predictions.
1059

1060 # SUPPLEMENTARY TABLES

1061 **Table S1. 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	24	NA[c]
In vitro growth (8h, 14h, 38h)	GSE43305	8	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	GSE60751	8	8[e]
Poor diet vs Rich diet	GSE60751	8	8[e]
Transition from exponential to stationary phase	GSE115054	16	NA[c]
codY KO	GSE23192	3	NA[c]
Sporulation TF deletions (*spo0A* KO, *sigEFGK* KOs, *spoIIID* KO)	GSE45977 GSE63777	18	6

1062 [a]Refers to the number of arrays used as numerator when estimating log2 ratios
1063 [b]Control arrays were averaged and used as denominator when estimating log2 ratios
1064 [c]Not Applicable. Dual channel array and therefore the control was included in each array
1065 [d]Six samples (half of the transcriptomes) in the GSE120189 dataset were used as controls to
1066 compute log2 ratios associated with the “Iron limitation” or “*fur* deletion” conditions
1067 [e]Eight samples (half of the transcriptomes) in the GSE60751 dataset were used as controls to
1068 either compute log2 ratios for the “Response to commensals” condition or the “Rich diet vs Poor
1069 diet” condition

1070 **Table S2. Compiled TF regulons**

Regulator	Regulon Size	Supporting Data	Reference
CcpA	194[a]	Protein-DNA binding, transcriptomics, *in silico*[b]	(Antunes et al., 2011, 2012)
CodY	160[c]	Protein-DNA binding, transcriptomics	(Dineen et al., 2010)
Fur	19	Transcriptomics, *in silico*[b]	(Berges et al., 2018; Dubois et al., 2016)
PrdR	181	Transcriptomics	(Bouillaut et al., 2019)
SigB	57	Transcriptomics	(Kint et al., 2017)
SigD	155	Transcriptomics	(El Meouche et al., 2013)
SigE	96	Transcriptomics	(Saujet et al., 2013)
SigF	25		
SigG	46		
SigH	40	Transcriptomics, *in silico*[b]	(Saujet et al., 2011)
SigK	54	Transcriptomics	(Saujet et al., 2013)
SigL	46[d]	Transcriptomics, *in silico*[b]	(Soutourina et al., 2020)
Spo0A	276[e]	Transcriptomics	(Fimlaid et al., 2013)

1071 [a]Only genes classified as CcpA-dependent (in the presence or absence of glucose) by Antunes
1072 et al. 2012 were included
1073 [b]*In silico* search of the binding motif of the corresponding regulator within the genes and their
1074 promoter sequences
1075 [c]Only genes negatively influenced by CodY were included
1076 [d]Only genes with SigL motif and downregulated in *sigL* deletion were included
1077 [e]Only genes positively influenced by Spo0A were included
1078
1079

1080 **Table S3. General properties of modules associated with the pathogenicity loci**

Gene	Locus tag[a]	Module[b]	Residual	Functional enrichment[c,d]	Enriched TF regulons[c,d]	Inferelator predicted regulators[a,c]
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, CD630_05790, PhoU
		395	0.51	NF	NF	CD_18100, CD_35440

1081 [a]The 'CD630_' segment of the locus tag was replace with 'CD_'
1082 [b]Modules that contain the corresponding member of the pathogenicity loci
1083 [c]Refers to the information shown for the corresponding module on the Cdiff Web Portal
1084 [d]'NF' indicates that no enrichment for functional terms or TF regulons was detected
1085

1086 **Table S4. Locus tag of uncharacterized genes associated with selected functional terms**
1087

Function Term	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
Multi-drug Transport Systems	29	NA	CD630_03790, CD630_03800, CD630_10991, CD630_33681
Aminobutanoate degradation	48	CodY, PrdR	CD630_05100, CD630_05101, CD630_08760, CD630_08780
Amino-acyl tRNA Synthesis and Modification	333	NA	CD630_12860, CD630_26940
Wood-Ljungdahl Pathway	159, 168	CcpA, GlnF	CD630_20210, CD630_29280
Xanthine, Hypoxanthine Transport and Metabolism	145	NA	CD630_17570, CD630_22580
Flagellar Motility	296	SigD	CD630_12920, CD630_13900
Ribosome Synthesis and Modification	317	NA	CD630_22121
Cobalamin Biosynthesis	157	NA	CD630_04760
Gram+ Peptidoglycan and Teichoic Acid	121	NA	CD630_15620
Chemotaxis	342	SigD	CD630_32720

1088
1089
1090

1091 **Table S5. Model predicted essential genes *in vivo***

1092

Locus tag	Gene name	Gene product/function	Pathway	*In silico* predictions		Experiment
				Essentiality *in vivo*	**Essentiality nutrient-rich broth *in vitro***	**Essentiality nutrient-rich broth *in vitro* (Tn-seq)**
CD630_23430	cat1	succinyl-CoA:coenzyme A transferase	Butanoate fermentation; Methionine Biosynthesis; Central carbon metabolism	Essential	Non-Essential	Non-Essential
CD630_34400	G12WB-3619	glycoside hydrolase-type carbohydrate-binding protein	Glycolysis	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_12230	drm	phosphopentomutase	Pentose-phosphate pathway; Nucleotide interconversion	Essential	Non-Essential	Non-Essential
CD630_15020	deoC	deoxyribose-phosphate aldolase	Pentose-phosphate pathway; Nucleotide interconversion	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_18160	cmk	cytidylate kinase	Pyrimidine metabolism	Essential	Non-Essential	Non-Essential

1093