RESEARCH ARTICLE   **Open Access**

CrossMark

Making life difficult for Clostridium difficile: augmenting the pathogen's metabolic model with transcriptomic and codon usage data for better therapeutic target characterization

Sara Saheb Kashaf[1]*, Claudio Angione[2] and Pietro Lió[1]

Abstract

Background: *Clostridium difficile* is a bacterium which can infect various animal species, including humans. Infection with this bacterium is a leading healthcare-associated illness. A better understanding of this organism and the relationship between its genotype and phenotype is essential to the search for an effective treatment. Genome-scale metabolic models contain all known biochemical reactions of a microorganism and can be used to investigate this relationship.

Results: We present *i*cdf834, an updated metabolic network of *C. difficile* that builds on *i*MLTC806cdf and features 1227 reactions, 834 genes, and 807 metabolites. We used this metabolic network to reconstruct the metabolic landscape of this bacterium. The standard metabolic model cannot account for changes in the bacterial metabolism in response to different environmental conditions. To account for this limitation, we also integrated transcriptomic data, which details the gene expression of the bacterium in a wide array of environments. Importantly, to bridge the gap between gene expression levels and protein abundance, we accounted for the synonymous codon usage bias of the bacterium in the model. To our knowledge, this is the first time codon usage has been quantified and integrated into a metabolic model. The metabolic fluxes were defined as a function of protein abundance. To determine potential therapeutic targets using the model, we conducted gene essentiality and metabolic pathway sensitivity analyses and calculated flux control coefficients. We obtained 92.3% accuracy in predicting gene essentiality when compared to experimental data for *C. difficile* R20291 (ribotype 027) homologs. We validated our context-specific metabolic models using sensitivity and robustness analyses and compared model predictions with literature on *C. difficile*. The model predicts interesting facets of the bacterium's metabolism, such as changes in the bacterium's growth in response to different environmental conditions.

Conclusions: After an extensive validation process, we used *i*cdf834 to obtain state-of-the-art predictions of therapeutic targets for *C. difficile*. We show how context-specific metabolic models augmented with codon usage information can be a beneficial resource for better understanding *C. difficile* and for identifying novel therapeutic targets. We remark that our approach can be applied to investigate and treat against other pathogens.

Keywords: Clostridium difficile, Metabolic networks, Metabolic pathways, Metabolic modeling, Genome scale modeling, Flux balance analysis, Sensitivity analysis, Antibiotic resistance

*Correspondence: ss2228@cam.ac.uk
[1]Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, CB3 0FD, Cambridge, UK
Full list of author information is available at the end of the article

Saheb Kashaf *et al. BMC Systems Biology* (2017) 11:25

Page 2 of 13

Background

Clostridium difficile is a gram-positive, spore-forming, anaerobic bacterium, which infects or colonizes various animal species. Clinical manifestations in humans range from asymptomatic colonization to mild diarrhea, pseudomembranous colitis, and death [1]. Infection by this bacterium is associated not only with significant patient morbidity and mortality, but also with a large economic burden for healthcare systems [2]. The primary risk factor for development of *C. difficile* infection among hospitalized patients is antibiotic use, which promotes toxicogenic *C. difficile* strains to proliferate, produce toxins, and induce disease [3]. Infection by this bacterium is most commonly associated with antibiotics such as clindamycin and amoxicillin [4]. Current recommendations for treatment of *C. difficile* infection (CDI) call for other antibiotics, such as metronidazole for mild infection cases and vancomycin for more severe cases [5]. The emergence of hypervirulent and antibiotic-resistant strains of this bacterium has motivated the search for novel methods of treating CDI. One method involves searching the bacterial central metabolic pathways for drug targets to create the next generation of antibiotics [6].

The quest to better understand this bacterium and identify novel drug targets against it can benefit vastly from a model of the genotype-phenotype relationship of its metabolism. Methods to model the genotype-phenotype relationship range from stochastic kinetic models [7] to statistical Bayesian networks [8, 9]. Kinetic models are limited as extensive experimental data is required to determine the rate laws and kinetic parameters of biochemical reactions. An alternative to kinetic models is metabolic modeling, which has been used to depict a range of cell types without the need for difficult-to-measure kinetic parameters [9]. Metabolic models have been able to predict cellular functions, such as cellular growth capabilities on various substrates, effect of gene knockouts at genome scale [10], and adaptation of bacteria to changes in their environment [11]. Metabolic models require a well-curated genome-scale metabolic network of the cell. Such networks contain all the known metabolic reactions in an organism, along with the genes that encode each enzyme involved in a reaction. The networks are constructed based on genome annotations, biochemical characterizations, and published literature on the target organism. The different scopes of such networks include metabolism, regulation, signaling, and other cellular processes [10].

Despite the success of metabolic modeling in capturing large-scale biochemical networks, the approach is limited as it describes cellular phenotype simply in terms of biochemical reaction rates and is thereby disconnected from other biological processes that impact phenotype.

Moreover, metabolic models cannot account for changes in the metabolism of the bacterium in response to different environmental conditions. Recent advances in the omic technologies, such as genomics (genes), transcriptomics (mRNA), and proteomics (proteins), have enabled quantitative monitoring of the abundance of biological molecules at various levels in a high-throughput manner. Integration of transcriptomic data has been shown to be effective in improving metabolic model predictions of cellular behavior in different environmental conditions [12].

Here we present an integrated model of the metabolism of *C.difficile* strain 630. We expanded the network *i*MLTC806cdf [13] with regards to various pathways, such as fatty acid, glycerolipid, and glycerophospholipid metabolism. Fatty acids are not only important components of bacterial cell membranes but they are also important intermediate metabolites in the production of vitamins, lipid A, and quorum sensing molecules [14]. The metabolism of phospholipids is also of interest as these compounds have been found to be closely tied to the growth phase in bacteria such as *Bacillus subtilis* [15, 16].

To bridge the gap between gene expression data and protein abundance, we accounted for the codon usage bias of the bacterium. During translation of a mRNA to a protein, the information contained in the form of nucleotide triplets (codons) in the RNA is decoded to derive the amino acid sequence of the resulting protein. Most amino acids are coded by two to six *synonymous codons*. These codons, which code for the same amino acid, are surprisingly used differentially in protein-encoding sequences [17]. The codon usage has been found to alter the translation time and the abundance of the resulting protein [18, 19]. To our knowledge, this is the first time codon usage has been quantified and incorporated into a genome-scale metabolic reconstruction.

We used the modified network and flux balance analysis [20] to simulate the steady-state metabolism of the bacterium. To understand the behavior of the bacterium in different environments, we integrated gene expression data. We incorporated the codon usage of the bacterium to bridge the gap between gene expression levels and protein abundance in the model. We then validated our metabolic models against the literature on the bacterium. Following this validation process, we used our models to identify potential drug targets. Essential genes have been previously proposed as potential therapeutic targets. [13]. We propose an additional method of predicting therapeutically-relevant genes through metabolic pathway sensitivity analysis and calculation of flux control coefficients. The choice of gene to target can be further refined by eliminating genes with a human homolog to reduce the off-target effects of the selected drug [13].

Saheb Kashaf *et al. BMC Systems Biology* (2017) 11:25

Page 3 of 13

Methods

Construction and validation of the metabolic model

icdf834: an expansion of the iMLTC806cdf network

In modifying the *i*MLTC806cdf network [13], we consulted KEGG [21] and incorporated some of the output from the review and curation of the MetaCyc [22] database for *C. difficile*, which was released on March 20, 2015. During curation, we manually considered the directionality and gene-reaction associations of each reaction in the existing network. We also manually expanded the existing network according to the procedure specified by Thiele et al. in [23]. We supported additions to the network with published literature on the bacterium. For example, the fatty acid profile found in *Clostridium difficile* is mostly dominated by C16:0, C16:1, C18:1, and C18:0 [24]. The major phospholipid types in this bacterium are phosphatidylglycerol analogs, with PG(31:2), PG(32:1), PG(33:2), PG(33:1) constituting the majority of these species [24]. Our modified network *icdf834* modifies and expands pathways concerning lipid metabolism in the existing network, such as those where compounds and reactions involved had been grouped together. By expanding the metabolism of the bacterium, we can also account for the wide array of fatty acids *C. difficile* can metabolize from its environment. This can provide important insights as many Gram-positive bacteria have been found to be able to incorporate and metabolize extracellular fatty acids [25]. When defining metabolic pathways in the expanded network, we used KEGG pathway identifiers so to remain consistent with the conventions employed in *i*MLTC806cdf [13].

The lipid component of the biomass equation of *i*MLTC806cdf had been obtained from the metabolic network of *Staphylococcus aureus* [26], where lipid compounds had been lumped together. There is a paucity of analyses on the chemical content of *C. difficile*'s biomass. Therefore, upon increasing the granularity of the network, we assumed coefficients from the biomass equation of the *i*Bsu1103 metabolic network developed for *Bacillus subtilis*, where these lumped lipid and teichoic acid species have been replaced by explicit species.

Constraint-based reconstruction and modeling approach

One constraint-based method for simulating the metabolic steady-state of a cell is flux-balance analysis (FBA), which can be used to analyze the metabolic network solely on the basis of systemic mass-balance and reaction capacity constraints. FBA simulations have been able to capture microorganism growth, nutritional resource consumption, and waste-product secretion rates of various cell types [27].

The first step of FBA involves representing the metabolic network in the form of a numerical matrix S of size $(m \times n)$. This matrix contains the stoichiometric coefficients of each of the m metabolites in the n different reactions. In the matrix, each row represents one unique metabolite and each column represents one reaction. The stoichiometric matrix helps enforce a mass balance constraint on the system. The mass balance on the cell for $i=1,\ldots,m$ metabolites and $j=1,\ldots,n$ reactions constrains the metabolite concentrations x_i, as shown in Eq. 1, where v_j is the flux through reaction j.

$$\frac{dx_i}{dt} = \sum_{j=1}^{n} S_{ij}v_j, i = 1, \ldots, m. \tag{1}$$

Under the steady state assumption $\frac{dx_i}{dt} = 0, \forall i$, the total amount of any compound being produced equals the total amount being consumed:

$$\sum_{j=1}^{n} S_{ij}v_j = 0, i = 1, \ldots, m. \tag{2}$$

In most metabolic models, there are more reactions than there are compounds [20]. Because there are more unknown variables than equations ($n>m$), any v that satisfies Eq. 2 is considered to be in the null space of S.

FBA can be used to find and determine points within the solution space that are most representative of the biological system using linear programming methods. Studies have revealed that metabolic fluxes in microorganisms are best predicted by maximizing the cellular objectives of growth [27]. To determine the point corresponding to the maximum growth rate within the constrained space, the objective function shown in Eq. 3 was maximized:

$$f(v) = c^T v, \tag{3}$$

where c is a vector of weights and indicates how much each reaction flux contributes to the biomass objective function. The maximum growth rate can be achieved by determining the flux distribution v that results in maximal biomass flux. Additional constraints can be added through the upper bound v_j^U and the lower bound v_j^L for the flux v_j. These bounds mandate the minimum and maximum fluxes allowed for a certain reaction and further decrease the space of allowable flux distributions for the relevant system. The mathematical representation of the metabolic reactions, the objective function, and the capacity constraints define a linear system as shown in Eq. 4.

$$
\begin{aligned}
\max \quad & c^T v \\
\text{subject to} \quad & Sv = 0 \\
& v_j^L \leq v_j \leq v_j^U, \ j = 1, \ldots, n.
\end{aligned}
\tag{4}
$$

The model fluxes are usually given units of $mmol/gDW \cdot h$, where gDW is the dry weight of cell mass in grams and h is the reaction time in hours. The bounds enforce thermodynamic constraints by dictating whether reactions are reversible or irreversible. The lower and upper flux

Saheb Kashaf *et al. BMC Systems Biology* (2017) 11:25

Page 4 of 13

bounds were arbitrarily chosen to be -10 $mmol/gDW \cdot h$ and 10 $mmol/gDW \cdot h$ for reversible reactions. For irreversible reactions, v_j^L was chosen to be 0 $mmol/gDW \cdot h$ and v_j^U was set to 10 $mmol/gDW \cdot h$. For our analysis, we used the COBRA toolbox 2.0 [28] in Matlab (version R2015b, Mathworks, Inc.).

Multi-objective optimization in metabolic models

One limitation of using only biomass as the objective is that goals in metabolism are often different and simultaneously competing so the scalar notion of "optimality" does not hold; examples of such trade-offs include maximizing energy production while minimizing protein costs [29]. Moreover, the biomass objective vector is usually perpendicular to one of the surfaces of the solution space of the FBA problem. Consequently, biomass maximizing flux states are usually degenerate; there exist multiple flux distributions that yield the same maximal biomass value [30]. To choose between the various flux distributions, additional criteria must be considered. For these reasons, we modeled metabolism as a multiobjective phenomenon. By modeling the metabolism of bacterium as a multi-objective problem, we address a conflict problem whereby maximizing one objective (eg. biomass) might involve a trade-off in the other objective (eg. intracellular flux); cells are thought to face a trade-off that is described by the set of Pareto-optimal solutions. We used a multi-objective optimization approach to address the z objectives, as shown in Eq. 5.

$$
\begin{aligned}
\max \quad & f(v) = (f_1(v), f_2(v), \ldots, f_z(v)) \\
\text{subject to} \quad & Sv = 0 \\
& v_j^L \leq v_j \leq v_j^U, \; j = 1, \ldots, n.
\end{aligned}
\tag{5}
$$

Note that, without loss of generality, we assumed that all the functions have to be maximized since minimizing a function $f(v)$ is equivalent to maximizing $-f(v)$.

Various works have attempted to systematically evaluate the ability of different objectives functions to reliably predict intracellular flux [31, 32]. According to their findings, bacterial metabolism can be better described by the objective of maximization of biomass or ATP production paired with the objective of minimization of intracellular flux [32]. Introducing the minimization of intracellular flux as a secondary objective allows for economic allocation of resources by the bacterium by selecting for metabolic routes that contain the fewest number of steps [33]. Thus, for our analyses we used maximization of biomass, along with minimization of intracellular flux as our objectives.

In a maximization multi-objective problem, a vector that is part of the feasible space is considered to be Pareto-optimal if all other vectors have the same or a lower value for at least one of the objective functions. Therefore, a Pareto-optimal solution is found when there exists no other feasible solution which would increase one objective without decreasing another objective. The set of Pareto-optimal solutions constitutes the Pareto-optimal front [34]. In the absence of additional information, no one Pareto-optimal solution can be said to be better than the other; higher-level information is required to choose one of the solutions [35].

As proposed by Costanza et al. [36], to solve this multi-objective optimization problem one can use bilevel linear programming coupled with evolutionary algorithms, namely stochastic optimization methods that simulate the process of natural evolution. Evolutionary algorithms are well suited to multi-objective problems because they can generate multiple Pareto-optimal solutions after one run and can use recombination to make use of the similarities of solutions [35]. The input to the evolutionary algorithm is a set of arrays, also called *individuals*, representing potential solutions to the problem. These arrays are then ranked based on the values of their objective functions. Potential optimal solutions are generated by retaining the best individuals and by generating new individuals through the use of variation. This process is continued until no further improvements are detected on the Pareto front. The population size and the number of populations used with this algorithm were 140 and 1400, respectively. To solve the linear programs, we used the Gurobi solver (v5.6.3, Gurobi Inc.) [37].

To validate our choice of objectives, we conducted a genetic analysis using multi-objective optimization. In this analysis, binary "knockout" vectors were created, with each containing a 1 in the location of a gene set to be off [36]. This analysis allowed us to determine how the growth of the organism changes in different environments, when genes may be turned on or off.

Robustness analysis

A facet of living organisms is their homeostasis, otherwise known as their ability to remain robust to external and internal perturbations within a certain range. External perturbations include changes in temperature or food supply while internal perturbations include spontaneous mutations. The robustness of biological systems is partly due to the presence of parallel metabolic pathways. Robustness represents the insensitivity of a system to changes in system parameters.

Global Robustness (GR) analysis can be used to survey the parameter space to determine the region where the cell exhibits specific features. More specifically, we perturbed the flux bounds of the metabolic model and observed the resulting effects on biomass production. The perturbation function $\gamma(\psi, \sigma)$ where γ applies noise σ, assumed to be Gaussian, to the system ψ for the trial τ. As proposed in [38], a robust trial is associated with a ρ of 1:

Saheb Kashaf *et al. BMC Systems Biology*   (2017) 11:25

Page 5 of 13

$$\rho(\psi, \tau, \phi, \epsilon) = \begin{cases} 1, & \text{if } |\phi(\psi) - \phi(\tau)| \le \epsilon \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where ϵ is the robustness threshold. The GR was defined as the percentage of trials determined to be robust. We arbitrarily defined ϵ to be 1% of the metric $\phi(\psi)$ and we arbitrarily limited the noise to 1%.

Incorporating transcriptomic and codon usage data in genome-scale models

To increase the reliability of the model, gene expression data was added to the FBA framework (Fig. 1). To relate this gene expression data to protein abundance, codon usage bias data was also incorporated. The translation rate of a codon is determined in part by the speed of diffusion of a translationally-competent tRNA to the ribosome. Because tRNAs are differentially abundant in the cell, codons pairing to high-abundance tRNAs are translated faster than those pairing to low-abundance tRNAs. Although synonymous codons produce the same amino acid sequence, they can alter the translation speed and the protein expression levels depending on the abundance of their associated tRNA [39]. Studies have revealed that a large codon bias generally resulted in higher protein expression levels [18, 19]. Therefore, the inclusion of codon bias can help improve the metabolic model predictions by helping link gene expression levels to protein levels.

The codon usage table for *C. difficile* was obtained from the Kazusa Codon Usage Database [40], which lists the frequency of different codons in the genome. The weights for synonymous codons was determined as the ratio between the observed frequency of the codon k and the frequency of the most preferred synonymous codon for that amino acid:

$$w_k = \frac{f_k}{\max(f_m)}, \text{where } k, m \in [\text{synonymous codons}]. \tag{7}$$

We obtained the mRNA sequence associated with the 834 genes of *C. difficile* from UniProt [41]. The counts of different codons were determined for each mRNA sequence. To obtain a measure of the codon bias, we calculated the Codon Adaptation Index (CAI) for each gene. The CAI represents the relative adaptiveness of the codon usage of the relevant gene to the codon usage of highly expressed genes [42]. The CAI ranges from 0 to 1, with a value of 1 indicating high expression and, by correlation, high abundance of the associated protein. The CAI represents the geometric mean of the weights corresponding to the codons in the sequence:

$$CAI = e^{\left[\frac{1}{L}\sum_{l=1}^{L}\ln(w_{k(l)})\right]}, \tag{8}$$
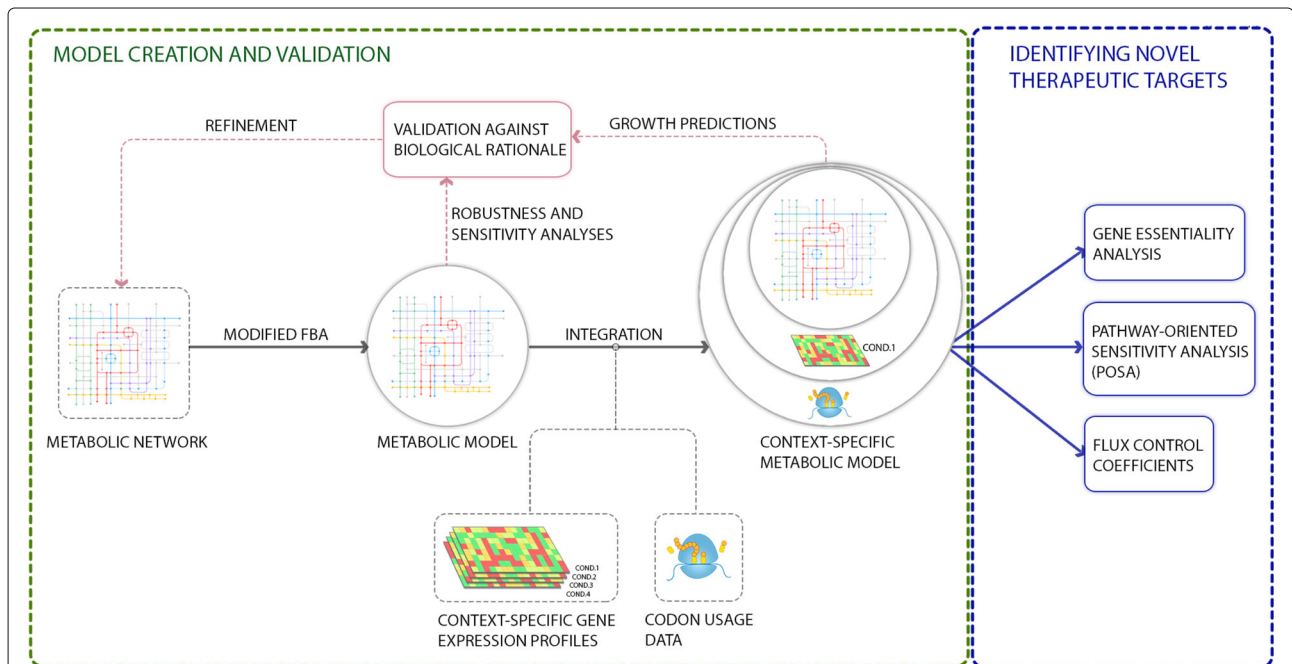
Fig. 1 Framework for modeling the metabolism of *C.difficile*. The updated metabolic network of the bacterium was used to create a metabolic model that was assessed using sensitivity and robustness analyses. Integrating gene expression and codon usage data yielded context-specific metabolic models that were evaluated against biological rationale and found fit for clinical applications. The augmented metabolic models were then used to identify potential therapeutic targets using gene essentiality analysis, PoSA, and flux control coefficient calculations

Saheb Kashaf *et al. BMC Systems Biology* (2017) 11:25

Page 6 of 13

where L is the number of codons in the genes and $w_{k(l)}$ is the weight associated with codon type k for lth codon along the length L of the gene. Because a large codon bias has been shown to result in higher protein expression levels, the gene expression data g_t for each gene t was scaled by *CAI* such that genes with the low codon bias had lower expression g'_t:

$$g'_t = g_t \cdot (CAI_t). \tag{9}$$

Each of the reactions in the metabolic model depends on a gene set, which is represented through the use of AND/OR operators. In this formulation, if a gene set is composed of two genes and an AND operator, both genes are required to carry out the corresponding reaction. On the other hand, if two genes connected by OR, one gene is sufficient in carrying out the reaction. This formulation can be transformed to derive the gene set expression GSE_j for gene set j of reaction j from the expression of individual genes g'_t, which in our case has been scaled by their respective codon usage. When two genes are connected through an AND operator, the gene set expression for reaction i, g_i, is the minimum of the scaled expression of the individual genes t making up the gene set. The gene set expression for two genes connected by an OR operator is the sum of the scaled expression of the individual genes. In each reaction of the model, to map the gene set expression into a specific condition of the model, we used the piecewise muliplicative function h and the associated h_j was adopted as a multiplicative factor for the flux bounds [43] :

$$v_j^L h(GSE_j) \leq v_j \leq v_j^U h(GSE_j),$$

where

$$h(GSE_j) = \begin{cases} (1 + |log(GSE_j)|)^{\frac{GSE_j - 1}{|GSE_j - 1|}} & \text{if } GSE_j \in \mathbb{R}^+ \setminus \{1\} \\ 1 & \text{if } GSE_j = 1 \end{cases} \tag{10}$$

The function h was chosen because at high mRNA abundance, an increase in mRNA abundance has been found to produce a relatively small increase in the protein synthesis rate. On the other hand, at low mRNA abundance, an increase in mRNA abundance has been found to produce a large increase in the protein synthesis rate [44].

Finally, we validated our context-specific metabolic models by incorporating codon usage and differential gene expression data into our model. We then compared trends in our models' biomass predictions to literature on the bacterium.

Prediction of therapeutic targets
Essential gene analysis
For each gene in the model, essential gene analysis involved removing reactions catalyzed by the gene or by a complex involving that gene and then using FBA [20] to predict growth. Genes were considered essential if following their removal, the predicted maximum growth rate was zero. The *C. difficile* R20291 (ribotype 027), for which gene essentiality data was available for comparison with our in silico results, had been grown on Tryptone-Glucose-Yeast Extract (TGY) broth. To approximate this medium, we used the complex medium defined by Larocque et al. during essential gene analysis of *i*MLTC806cdf [13].

Pathway-oriented sensitivity analysis
The growing research attention on metabolic pathways, rather than on specific reactions, is motivated by novel methods that allow for a better understanding of the functionality of complex webs of metabolic reactions. To date, much of the study of metabolic pathways, their crosstalks, and their role in the overall metabotype has been carried out with statistical and model-based approaches [45, 46].

Sensitivity analysis is used to identify model inputs that have a large influence on the model outputs. To find the metabolic pathways that have the largest effect on the outputs of *i*MLTC806cdf and *i*cdf834, we used Pathway-oriented Sensitivity Analysis (PoSA) [36]. PoSA involves genetically manipulating the metabolic model to find the *sensitive* pathways, which make a large impact on model outputs. In other words, we perturbed pathways by mutating the genes that govern their biochemical reactions and analyzed the result on the outputs. In the knockout vector $y = \{b_1, b_2, \ldots, b_s, \ldots, b_p\}$, b_s represents the perturbations on the genes governing the metabolic pathway s, where $|b_s| = W_s$ (number of genes partaking in the sth pathway). Because the gene knockouts are represented through the use of binary variables, we perform combinatorial perturbations, namely the bits in b_s are switched from 0 to 1 or from 1 to 0; note that if a gene in b_s is set to 1, this gene is knocked-out in the model.

According to [36], the Pathway Elementary Effect (PEE) for the genetic perturbation b_s can be defined as follows:

$$PEE_s = \frac{\| F(b_1, b_2, \ldots, \tilde{b}_s, \ldots, b_p) - F(\tilde{y}) \|}{\Delta_s}, \tag{11}$$

where $\tilde{b}_s$ represents the genetic manipulation of the input b_s; $\tilde{y}$ is the mutation carried out on the knockout vector y; $F(y)$ is the vector v of fluxes as produced by the model; finally, Δ_s is a scale factor defined as:

$$\Delta_s = \frac{1}{W_s} \sum_{i=1}^{W_s} \tilde{b}_s(i), \quad s = 1, \ldots, p. \tag{12}$$

Next, the sensitivity indices μ and σ are determined by calculating the mean and the standard deviation of the distribution of the PEE for each input. Pathways with a large μ have a large influence on the output. A large σ

Saheb Kashaf *et al. BMC Systems Biology* (2017) 11:25

Page 7 of 13

indicates an input whose influence highly depends on the value of other inputs. By perturbing the genes through the use of knockouts and comparing the outputs of the model with and without the genetic manipulations, we detected the most sensitive pathways of the metabolic models.

Calculation of flux control coefficients

PoSA provides valuable information on sensitive pathways that can be targeted by therapies, but often more specific drug target predictions are desired. To understand how a metabolic pathway is controlled and can be altered, its control structure has to be determined. The *flux control coefficient* [47] is the flux v_{ydh} through a particular reaction, catalyzed by enzyme *ydh*, of the metabolic pathway with respect to the concentration x_{xase} of an enzyme xase:

$$C_{x_{xase}}^{v_{ydh}} = \frac{\partial v_{ydh}}{\partial x_{xase}} \cdot \frac{x_{xase}}{v_{ydh}} = \frac{\partial \ln v_{ydh}}{\partial \ln x_{xase}} \qquad (13)$$

In our calculations, the enzyme concentration was assumed to be equal to the gene expression level adjusted by CAI. When calculating the flux control coefficients, we considered a 1% perturbation in the enzyme concentration x_{xase}. Flux control coefficients provide a quantitative measure of the degree of control an enzyme exerts on a metabolite flux and can quantitatively substitute for the qualitative concept of essential gene [48]. Thus, they can be used to identify steps that should be modified to achieve a successful alteration of the flux in outputs of clinical (e.g. drug therapy) relevance.

Analysis of cDNA microarrays

We used microarray analysis to determine the combination of genes which were up-regulated or down-regulated in different environmental conditions. We used *Limma* [49], a package in Bioconductor 3.1, for statistical analysis of gene expression. We preprocessed the data through background correction, within-array normalization, and between-array normalization. After normalization, we used filtering to remove probes that did not appear to be expressed in any of the experimental conditions. Next, we used linear models to analyze the microarray data. To conduct statistical analysis and assess differential expression, we used an empirical Bayes method to modulate the standard errors of the log-fold changes. To test for the comparisons of interest, we used an analysis of variance (ANOVA) model.

Results and discussion

Expansion and modification of *i*MLT806cdf to *i*cdf834

The genome of *C. difficile* strain 630 is composed of a circular chromosome of 4,290,252 bp coding for 3968

open reading frames (ORFs), along with a plasmid containing 7881 bp coding for 11 ORFs [50]. The modified metabolic network draft contains 21% of the ORFs present in the chromosomal genome of the bacteria with 834 ORFs, a modest improvement upon *i*MLTC806cdf, which contains 806 ORFs, as shown in Table 1. Our expanded metabolic network also consists of 807 metabolites and 1227 reactions. The final version of the network is available as an SBML file and as an Excel file that indicate the reactions, metabolites, genes, and compartments involved in the metabolic network, along with references to literature that support additions or modifications to the existing network. The new network has two additional dead-end metabolites as compared with those found in *i*MLTC806cdf. The Excel and SBML file, along with the the justification for keeping the dead-end metabolites in the model, have been uploaded to http://github.com/ssahebkashaf/Peptoclostridiumdifficile630. The code for all of the analyses employed in our work is also freely available on this repository.

We repeated analyses previously conducted by Larocque et al. to validate *i*MLTC806cdf [see Additional file 1]. Namely, we compared the ability of *i*cdf834 and *i*MLTC806cdf to identify essential amino acids and metabolizable carbon sources. The removal of amino acids that were not found to be essential or to affect growth, did not affect model-predicted biomass production in both models. Moreover, no biomass was produced in the absence of essential amino acids (cysteine, leucine, isoleucine, proline, tryptophan and valine) [51] in both models. Therefore, similar to the previous network, our network is able to account for the essentiality of various amino acids on the growth of *C.difficile*.

With regards to carbon sources, both models were able to correctly predict a range of carbon sources that are utilized by the bacterium. Moreover, the bacterium was able

Table 1 Comparison of the metabolic network *iMLTC806cdf* published by [13] and the modified and expanded network *i*cdf834

Features		Number	
Genomic Information of C. difficile			
	Genome size (bp)	4,290,252	
	Open reading frames	3968	
Reconstructed models			
		iMLTC806cdf	*icdf834*
	Metabolites	703	807
	Reactions	1091	1227
	Open reading frames	806	834

Saheb Kashaf *et al. BMC Systems Biology* (2017) 11:25

Page 8 of 13

to generate biomass in the absence of other carbon substrates, such as fructose, mannose, mannitol, and sorbitol. This finding is consistent with literature, which maintains that *C. difficile* is not restricted to metabolizing sugars and can ferment other compounds, even amino acids, to obtain both its carbon and energy [52].

Validation of metabolic models
Genetic analysis using multi-objective optimization
Our modeling approach is intended to simulate the conflicting objectives faced by the bacterium, where optimal performance in one objective coincides with sub-optimal performance in another objective. We used a knockout parameter space to find the genetic designs that would optimize the two objectives. In Fig. 2, we show the areas of objective space discovered by the genetic algorithm during the genetic analysis from the first generation to generation 1400. The optimization algorithm adaptively moves to regions that maximize biomass while minimizing the total intracellular flux, as evident in the curvature of the plot in Fig. 2. After conducting the genetic analysis, the Pareto front, shown in black in the inset of Fig. 2, was determined. The Pareto front is the set of nondominated solutions that represents the range of phenotypes resulting from different trade-offs between the two objectives. The presence of a Pareto front, as opposed to a singular dominated solution, aligned with our a priori expectations regarding the metabolic plasticity inherent to the bacterium [53]. Our findings, along

Fig. 2 Genetic analysis using multi-objective optimization. Regions of objective space explored by the optimization algorithm for the objectives of maximization of biomass and minimization of total intracellular flux. Solutions are represented by progressively warmer colors depending on the time step of the algorithm in which they had been adaptively generated from the initial point. The Pareto front is shown in *black* in the inset

with previous literature on the choice of objectives, supported our choice of objectives to model *C. difficile*'s metabolism.

Robustness analysis
We gauged the robustness of our model by determining the change in the maximal biomass flux in response to different perturbations. Global Robustness (GR) analysis revealed that the biomass production was fully robust to perturbations for a flux bound perturbation (σ) and a tolerance (ϵ) of 1% [see Additional file 2]. The GR falls when σ is increased or ϵ is decreased. This facet of the bacterium's metabolism was biologically relevant as bacteria such as *C. difficile* are able to grow despite small fluctuations in their physical environment. Robustness analysis illustrated that the global behavior of our metabolic model matches our expectations from biological rationale and supported the use of our models to predict the behavior of the bacterium in different environments.

Changes in C. difficile's growth in different conditions
We obtained the relevant microarray datasets from the Gene Expression Omnibus (GEO) database [54] under the accession numbers GSE22423 and from the ArrayExpress database [55] under the accession numbers E-GEOD-37442 and E-BUGS-56. Context-specific models for *C. difficile* were generated by incorporating gene expression data obtained for the bacterium in different environmental conditions. To improve the reliability of the model, we also integrated codon usage data. Model predictions of these context-specific models were compared to expectations about the organism's behavior from literature.

Previous work suggests that sub-MIC concentrations of amoxicillin, metronidazole, and clindamycin slowed growth of toxigenic *C. difficile* as compared with the controls [56]. To test these findings in silico, we incorporated gene expression levels of *C. difficile* in response to sub-MIC levels of different antibiotics into our model. As compared with the *C. difficile* grown on BHI broth, toxigenic strains of *C. difficile* grown on sub-inhibitory concentrations of antibiotics exhibited reductions in their biomass, with those grown on amoxicillin showing the smallest growth (as shown in Table 2). This finding is supported by literature [57] that has shown that in vitro, amoxicillin is effective against *C. difficile*. These findings have lead to speculations that in vivo, this antibiotic is effective agisnt vegetative forms of the bacterium but not against *C. difficile* spores [58]. Another potential explanation is that this broad-spectrum antibiotic may impair the intestinal microflora in a way that supports proliferation of *C. difficile*.

Additionally, the decline in biomass production following heat shock from 30 to 43 °C shown in Table 2 could be due to the general stress response employed by the

Saheb Kashaf *et al. BMC Systems Biology* (2017) 11:25

Page 9 of 13
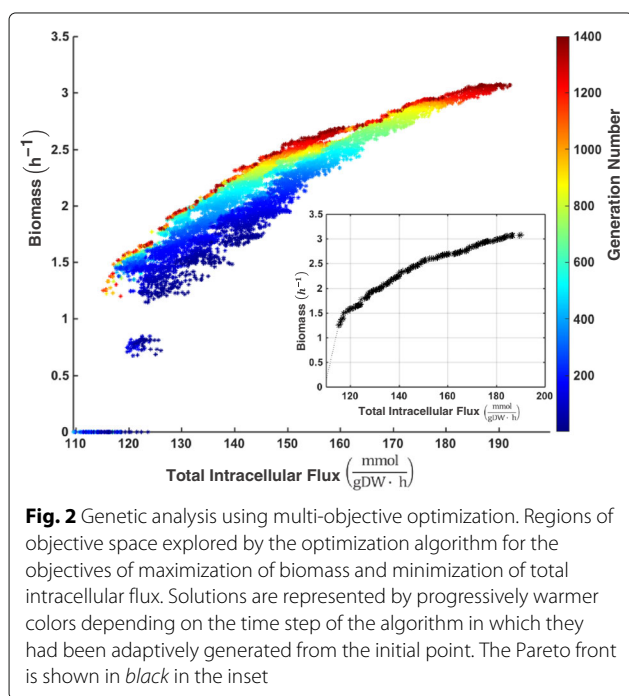
Table 2 Percent change in model-predicted biomass production (growth) of *C. difficile* in different conditions

Microarray data accession number/database	Condition	% change in biomass (h^{-1})
E-GEOD-37442/ ArrayExpress		
	Heat shock from 30 °C to 43 °C	↓ 24.3%
E-BUGS-56/ ArrayExpress		
	Sub-MIC level of amoxicillin	↓ 27.4%
	Sub-MIC level of clindamycin	↓ 16.6%
	Sub-MIC level of metronidazole	↓ 2.3%
	BHI broth	↑ 1.0%
GSE22423/GEO		
	Supplementation of 10mM cysteine	↑ 1.1%

The microarray data for each condition was obtained from the GEO or ArrayExpress databases, using the specified accession numbers. The differential gene expression levels obtained from analysis of this microarray data was used to make a metabolic model for each condition. These context-specific metabolic models were used to predict change in biomass production for each condition compared with the control of each microarray dataset

bacterium. The heat shock response of *C. difficile* has been found to be involve gene clusters homologous to *E. coli* heat-shock operons [59]. The heat shock response in *E. coli* has been found to be associated with a decrease in central carbon metabolism and a decline in cellular growth [60]. Literature on related bacteria is thereby in agreement with the model's prediction of a significant reduction in growth in *C. difficile* following the heat shock. Additionally, according to the work of Dubois et al., the supplementation of 10mM cysteine to the medium did not affect *C. difficile*'s growth [61]. After integrating the microarray data from their work, we found that our in silico findings agreed with their experimental results.

Validation of the findings of our context-specific metabolic models against the literature on the bacterium showed that metabolic models allow for an enriched view of omic data and may be valuable tools for better understanding the behavior of *C. difficile* in different conditions.

## Prediction of therapeutic targets
Gene essentiality analysis
Essential genes have been cited as promising targets for development of new antimicrobials due to their importance for bacterial survival [62]. Using FBA, we performed

an in silico gene deletion study to predict potential essential genes that may lead to the identification of new drug targets. This analysis had already been conducted for *i*MLTC806cdf based on a 5% threshold, and gene essentiality results had been compared to genes deemed essential for *B. subtilis*, for which this data had been available [13]. We performed gene essentiality analysis for both *i*MLTC806cdf and *i*cdf834 and validated our results using recently available literature on the essential genes of the *C. difficile* R20291 (ribotype 027) [63]. While *i*MLTC806cdf predicted 48 essential genes and had a 86.5% accuracy in predicting gene essentiality, *i*cdf834 predicted 46 essential genes and had a 92.3% accuracy [see Additional file 3].

Pathway-oriented sensitivity analysis and flux control coefficients
For our PoSA analysis, we chose the gene expression profile of the bacterium when grown on BHI broth. Each pathway was assessed through random perturbations of its reactions, and the average perturbation μ and the standard deviation σ were computed as a result. We performed the pathway-based sensitivity analysis and identified sensitive pathways before and after modifying the metabolic model as shown in Fig. 3. The pathway with the largest μ, and thereby the greatest control on biomass production or growth in both *i*MLTC806cdf and *i*cdf834 is the valine, leucine, and isoleucine metabolism pathway. These three amino acids are essential to the bacterium and their metabolism was also expected to be essential. The second most sensitive pathway is alanine, aspartate, and glutmate metabolism in *i*MLTC806cdf and glycolysis/gluconeogenesis in *i*cdf834. Additional sensitive pathways in *i*cdf834 include pyrimidine metabolism and pyruvate metabolism. Model findings suggest that therapies against infection may likely be more effective if they target key enzymes in these sensitive pathways.

To find more specific therapeutic targets, flux control coefficients for enzymes on biomass production in the metabolic model were determined and compared for BHI broth (E-BUGS-56), cysteine supplementation (GSE22423), and heat shock (E-GEOD-37442) gene expression data. The four enzymes with largest flux control coefficients in each condition are shown in Fig. 4, while the complete list of flux control coefficients in different conditions has been uploaded to the public repository. These flux control coefficients were interestingly involved in pathways deemed sensitive during PoSA. These enzymes varied amongst the four conditions, suggesting that access to the in vivo gene expression profile of *C. difficile* can be used to predict better drug targets for patients. Therapies aimed at reducing growth of *C. difficile* should target enzymes with high flux

Saheb Kashaf *et al. BMC Systems Biology* (2017) 11:25

Page 10 of 13
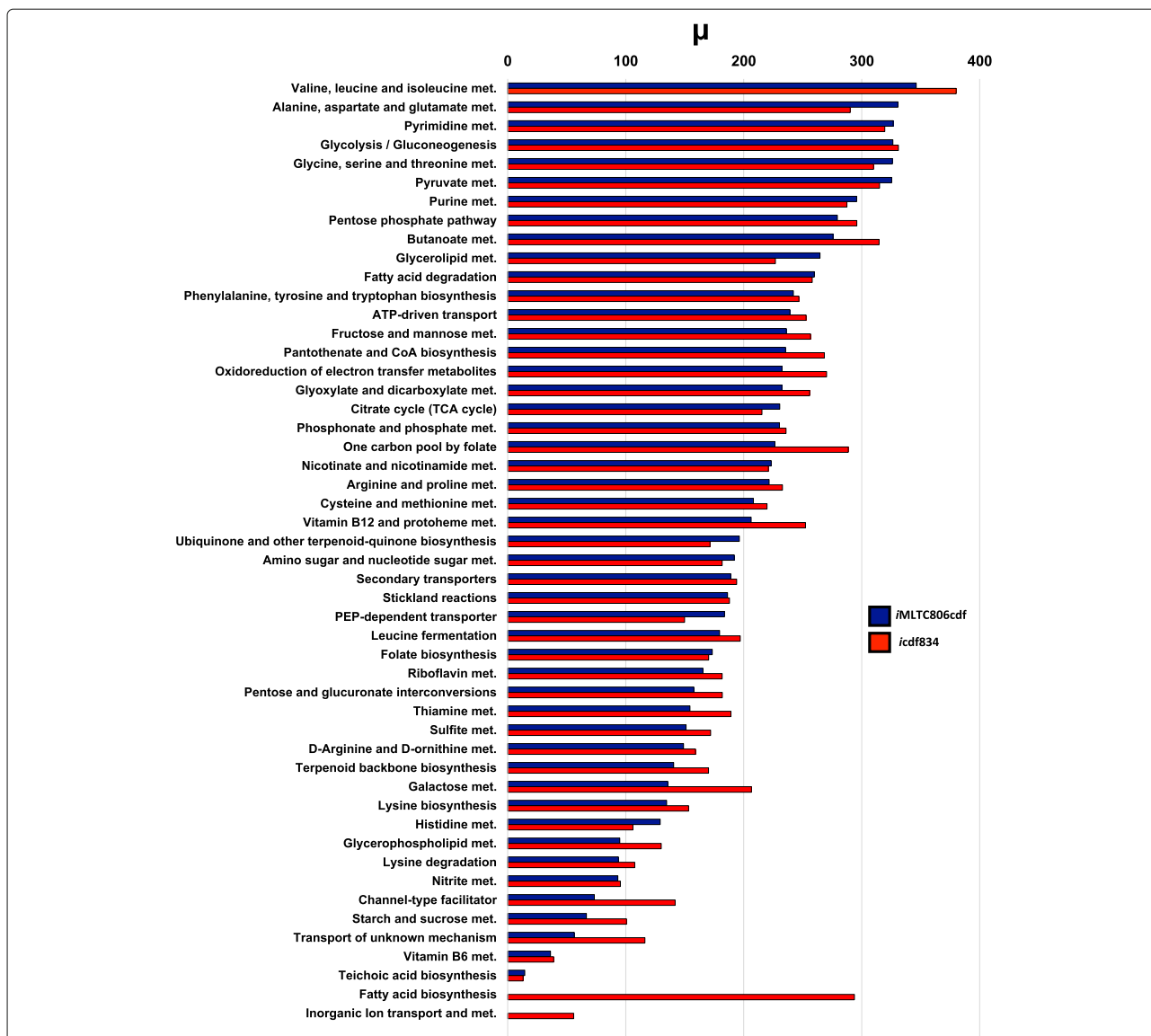
Fig. 3 PoSA was used to compare the most sensitive pathways of *i*MLTC806cdf and *i*cdf834. The *i*MLTC806cdf model is composed of 48 metabolic pathways and the *i*cdf834 model is composed of 50 metabolic pathways. Biomass production is most sensitive to pathways with higher calculated μ
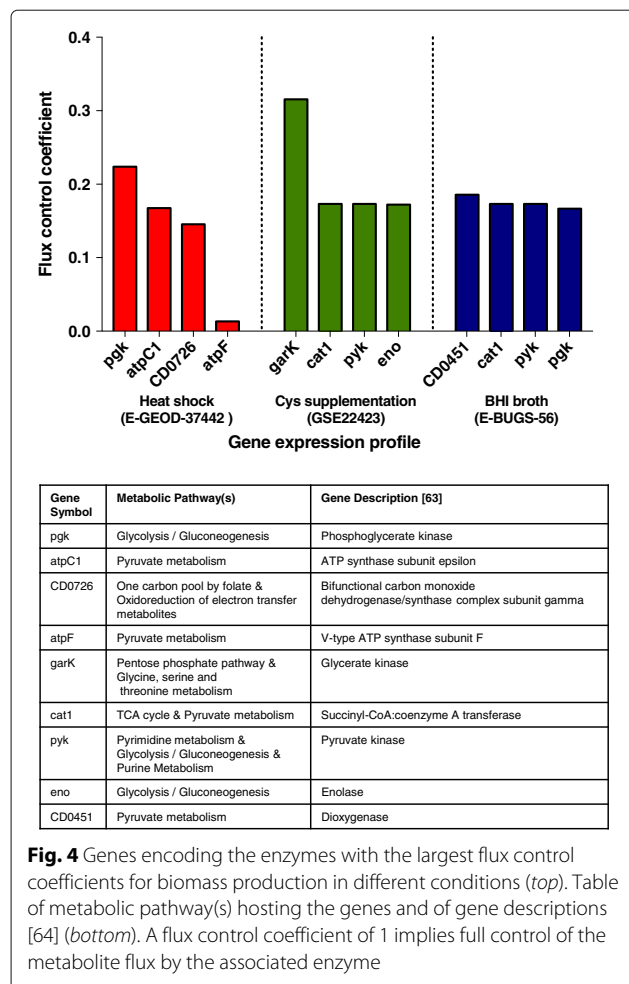
coefficients as, according to our model, their activity is most closely tied to biomass production.

Conclusion

In this study, we expanded the existing metabolic network for *C. difficile* and used it to create context-specific metabolic models of its metabolism that allow us to understand how the bacterium alters its metabolism depending on its environment. To predict the bacterium's behavior in different environmental conditions, the model was integrated with transcriptomic and codon usage data to generate reliable and context-specific metabolic flux distributions. We validated the model by conducting robustness and sensitivity analyses. We

further assessed its predictive potential by comparing model predictions with published experimental data to gauge the consistency of model findings with the current knowledge of *C. difficile*'s metabolism. Through this literature-based validation, we found that the model is a valuable tool for qualitatively understanding the behavior of the bacterium in different settings. The model can also be used to find potential therapeutic targets by allowing for determination of essential genes and context-specific sensitive pathways and flux control coefficients.

Context-specific metabolic models can allow for a better understanding different medically-relevant conditions (eg. pre-infection, post infection) and can be continuously

Saheb Kashaf *et al. BMC Systems Biology*   (2017) 11:25

Page 11 of 13

Gene Symbol	Metabolic Pathway(s)	Gene Description [63]
pgk	Glycolysis / Gluconeogenesis	Phosphoglycerate kinase
atpC1	Pyruvate metabolism	ATP synthase subunit epsilon
CD0726	One carbon pool by folate & Oxidoreduction of electron transfer metabolites	Bifunctional carbon monoxide dehydrogenase/synthase complex subunit gamma
atpF	Pyruvate metabolism	V-type ATP synthase subunit F
garK	Pentose phosphate pathway & Glycine, serine and threonine metabolism	Glycerate kinase
cat1	TCA cycle & Pyruvate metabolism	Succinyl-CoA:coenzyme A transferase
pyk	Pyrimidine metabolism & Glycolysis / Gluconeogenesis & Purine Metabolism	Pyruvate kinase
eno	Glycolysis / Gluconeogenesis	Enolase
CD0451	Pyruvate metabolism	Dioxygenase

Fig. 4 Genes encoding the enzymes with the largest flux control coefficients for biomass production in different conditions (*top*). Table of metabolic pathway(s) hosting the genes and of gene descriptions [64] (*bottom*). A flux control coefficient of 1 implies full control of the metabolite flux by the associated enzyme

refined by integrating novel information regarding *C. difficile*'s metabolism. Our model can be used by biomedical researchers to study the bacterium and devise targeted treatments. Our approach can also be scaled-up to simulate the interactions between the gut microbiota and the host using a bottom-up modeling approach. By accounting for the gut microbiota-host interactions, we can construct a whole gut model response to infections and other inflammatory events, paving the path towards more informed and effective treatments.

Availability and requirements
Project name: *i*cdf834
Project home page: http://github.com/ssahebkashaf/Peptoclostridiumdifficile630
Operating system: platform independent
Programming language: MATLAB
Other requirements: Gurobi
License: University of Cambridge
Any restrictions to use by non-academics: license needed

Additional files

Additional file 1: Carbon source usage, carboxylic acids secreted, and essential amino acids in *i*cdf834 as compared with *i*MLTC806cdf and literature. (PDF 41 kb)

Additional file 2: GR of biomass production in *i*cdf834 to internal and external perturbations. (PDF 656 kb)

Additional file 3: Gene essentiality analysis of *i*cdf834 and *i*MLTC806cdf. (PDF 49 kb)

Abbreviations
CAI: Codon adaptation index; FBA: Flux balance analysis; GR: Global robustness; ORF: Open reading frame; PEE: Pathway elementary effect; PoSA: Pathway-oriented sensitivity analysis; TGY: Tryptone-glucose-yeast extract

Availability of data and materials
The metabolic network supporting the conclusions of this article, along with all the code used in the analyses, is freely available at http://github.com/ssahebkashaf/Peptoclostridiumdifficile630.

Authors' contributions
SS, CA, and PL conceived the study, designed and developed the methodology. SS, CA, and PL wrote the code and conducted the simulations and analysis. SS, CA, and PL coordinated the study, collected the data, and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, CB3 0FD, Cambridge, UK. [2]Department of Computer Science and Information Systems, Teesside University, Borough road, TS1 3BA, Middlesbrough, UK.

References
1. Trudel JL. Clostridium difficile colitis. Clin Colon Rectal Surg. 2007;20(1): 13–7. doi:10.1055/s-2007-970195.
2. Dubberke E. Clostridium difficile infection: the scope of the problem. J Hosp Med. 2012;7 Suppl 3(March):1–4. doi:10.1002/jhm.1916.
3. Janvilisri T, Scaria J, Thompson AD, Nicholson A, Limbago BM, Arroyo LG, Songer JG, Gröhn YT, Chang YF. Microarray identification of Clostridium difficile core components and divergent regions associated with host origin. J Bacteriol. 2009;191(12):3881–91. doi:10.1128/JB.00222-09.
4. Mylonakis E, Ryan ET, Calderwood SB. Clostridium difficile–associated diarrhea: a review. Arch Intern Med. 2001;161(4):525–33. doi:10.1001/archinte.161.4.525.
5. Gerding DN, File TM, McDonald LC. Diagnosis and treatment of Clostridium difficile Infection. Infect Dis Clin Prac. 2016;24(1):3–10. doi:10.1097/IPC.0000000000000350.

Saheb Kashaf *et al. BMC Systems Biology* (2017) 11:25

Page 12 of 13

6. Murima P, McKinney JD, Pethe K. Targeting bacterial central metabolism for drug development. Chem Biol. 2014;21(11):1423–32. doi:10.1016/j.chembiol.2014.08.020.

7. Wilkinson DJ. Stochastic modelling for quantitative description of heterogeneous biological systems. Nat Rev Genet. 2009;10(2):122–33. doi:10.1038/nrg2509.

8. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nat Rev Genet. 2009;10(10):681–90. doi:10.1038/nrg2615.

9. Bordbar A, Monk JM, King ZA, Palsson BØ. Constraint-based models predict metabolic and associated cellular functions,. Nat Rev Genet. 2014;15(2):107–20. doi:10.1038/nrg3643.

10. O'Brien EJ, Monk JM, Palsson BØ. Using genome-scale models to predict biological capabilities. Cell. 2015;161(5):971–87. doi:10.1016/j.cell.2015.05.019.

11. Angione C, Lió P. Predictive analytics of environmental adaptability in multi-omic network models. Sci Rep. 2015;5:15147. doi:10.1038/srep15147.

12. Zelezniak A, Sheridan S, Patil KR. Contribution of network connectivity in determining the relationship between gene expression and metabolite concentration changes. PLoS Comput Biol. 2014;10(4). doi:10.1371/journal.pcbi.1003572.

13. Larocque M, Chénard T, Najmanovich R. A curated C. difficile strain 630 metabolic network: prediction of essential targets and inhibitors. BMC Syst Biol. 2014;8(1):117. doi:10.1186/s12918-014-0117-z.

14. Cheng J, Ma J, Lin J, Fan ZC, Cronan JE, Wang H. Only one of the five Ralstonia solanacearum long-chain 3-ketoacyl-acyl carrier protein synthase homologues functions in fatty acid synthesis. Appl Environ Microbiol. 2012;78(5):1563–73. doi:10.1128/AEM.07335-11.

15. Lobasso S, Palese LL, Angelini R, Corcelli A. Relationship between cardiolipin metabolism and oxygen availability in Bacillus subtilis. FEBS Open Bio. 2013;3:151–5. doi:10.1016/j.fob.2013.02.002.

16. Den Kamp JA, Redai I, van Deenen LL. Phospholipid composition of Bacillus subtilis. J Bacteriol. 1969;99(1):298–303.

17. Ermolaeva MD. Synonymous codon usage in bacteria. Curr Issues Mol Biol. 2001;3(4):91–7.

18. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI. A sampling of the yeast proteome. Mol Cell Biol. 1999;19(11):7357–368. doi:10.1128/MCB.19.11.7357.

19. Klumpp S, Dong J, Hwa T. On ribosome load, codon bias and protein abundance. PLoS ONE. 2012;7(11). doi:10.1371/journal.pone.0048542.

20. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010;28(3):245–8. doi:10.1038/nbt.1614. http://arxiv.org/abs/NIHMS150003NIHMS150003.

21. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44(D1):457–62. doi:10.1093/nar/gkv1070.

22. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Karp PD. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2016;44(D1):471–80. doi:10.1093/nar/gkv1164.

23. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 2010;5(1):93–121. doi:10.1038/nprot.2009.203. http://arxiv.org/abs/NIHMS150003NIHMS150003.

24. Drucker DB, Wardle HM, Boote V. Phospholipid profiles of Clostridium difficile. J Bacteriol. 1996;178(19):5844–6.

25. Parsons JB, Rock CO. Is bacterial fatty acid synthesis a valid target for antibacterial drug discovery? Curr Opinion Microbiol. 2011;14(5):544–9. doi:10.1016/j.mib.2011.07.029.

26. Heinemann M, Kümmel A, Ruinatscha R, Panke S. In silico genome-scale reconstruction and validation of the Staphylococcus aureus metabolic network. Biotechnol Bioeng. 2005;92(7):850–64. doi:10.1002/bit.20663.

27. Ibarra RU, Edwards JS, Palsson BØ. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature. 2002;420(6912):186–9. doi:10.1038/nature01149.

28. Schellenberger J, Que R, Fleming RM, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, et al. Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox v2. 0. Nat Protoc. 2011;6(9):1290–307. doi:10.1038/nprot.2011.308.

29. Molenaar D, van Berlo R, de Ridder D, Teusink B. Shifts in growth strategies reflect tradeoffs in cellular economics. Mol Syst Biol. 2009;5(323):323. doi:10.1038/msb.2009.82.

30. Lee JS, Nishikawa T, Motter AE. Why optimal states recruit fewer reactions in metabolic networks. Discrete Continuous Dynamical Syst. 2012;32(8):2937–50. doi:10.3934/dcds.2012.32.2937. http://arxiv.org/abs/1206.07661206.0766.

31. Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol Syst Biol. 2007;3:119. doi:10.1038/msb4100162.

32. García Sánchez CE, Torres Sáez RG. Comparison and analysis of objective functions in flux balance analysis. Biotechnol Prog. 2014;30(5):985–91. doi:10.1002/btpr.1949.

33. Sweetlove LJ, Ratcliffe RG. Flux-balance modeling of plant metabolism. Front Plant Sci. 2011;2(August):1–10. doi:10.3389/fpls.2011.00038.

34. Bui LT, Alam S. An introduction to multi-objective optimization. In: Multi-objective Optimization in Computational Intelligence: Theory and Practice; 2008. p. 1–19. doi:10.4018/978-1-59904-498-9.

35. Sbalzarini IF, Muller S, Koumoutsakos P. Multiobjective optimization using evolutionary algorithms. In: Proceedings of the Summer Program. 2000. p. 63–74.

36. Costanza J, Carapezza G, Angione C, Lió P, Nicosia G. Robust design of microbial strains. Bioinformatics. 2012;28(23):3097–104. doi:10.1093/bioinformatics/bts590.

37. Gurobi Optimization I. Gurobi Optimizer Reference Manual. 2016. http://www.gurobi.com. Accessed 20 Feb 2016.

38. Angione C, Costanza J, Carapezza G, Lió P, Nicosia G. Multi-target analysis and design of mitochondrial metabolism. PLoS ONE. 2015;10(9). doi:10.1371/journal.pone.0133825.

39. Gorochowski TE, Ignatova Z, Bovenberg RAL, Roubos JA. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. Nucleic Acids Res. 20151–11. doi:10.1093/nar/gkv199.

40. Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from the international DNA sequence databases; its status 1999. Nucleic Acids Res. 1999;27(1):292. doi:10.1093/nar/27.1.292.

41. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LSL. The universal protein resource (UniProt). Nucleic Acids Res. 2005;33. doi:10.1093/nar/gki070.

42. Lee S, Weon S, Lee S, Kang C. Relative codon adaptation index, a sensitive measure of codon usage bias. Evol Bioinforma. 2010;2010(6):47–55. doi:10.4137/EBO.S4608.

43. Angione C, Conway M, Lió P. Multiplex methods provide effective integration of multi-omic data in genome-scale models. BMC Bioinforma. 2016;17(S4):83. doi:10.1186/s12859-016-0912-1.

44. Firczuk H, Kannambath S, Pahle J, Claydon A, Beynon R, Duncan J, Westerhoff H, Mendes P, McCarthy JE. An in vivo control map for the eukaryotic mRNA translation machinery. Mol Syst Biol. 2013;9(635):635. doi:10.1038/msb.2012.73.

45. Angione C, Pratanwanich N, Lió P. A hybrid of metabolic flux analysis and bayesian factor modeling for multiomic temporal pathway activation. ACS Synthetic Biol. 2015. doi:10.1021/sb5003407.

46. Schilling CH, Edwards JS, Letscher D, Palsson BØ, et al. Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. Biotechnol Bioeng. 2000;71(4):286–306.

47. Mazat JP, Reder C, Letellier T. Why are most flux control coefficients so small? J Theor Biol. 1996;182(3):253–8. doi:10.1006/jtbi.1996.0162.

48. Cascante M, Boros LG, Comin-Anduix B, de Atauri P, Centelles JJ, Lee PW-N. Metabolic control analysis in drug discovery and disease. Nat Biotechnol. 2002;20(3):243–9. doi:10.1038/nbt0302-243.

49. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Nucleic Acids Res1–13. doi:10.1093/nar/gkv007.

50. Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson NR, Roberts AP, Cerdeño-Tárraga AM, Wang H, Holden MTG, Wright A, Churcher C, Quail MA, Baker S, Bason N, Brooks K, Chillingworth T, Cronin A, Davis P, Dowd L, Fraser A, Feltwell T, Hance Z, Holroyd S, Jagels K, Moule S, Mungall K, Price C, Rabbinowitsch E, Sharp S, Simmonds M, Stevens K, Unwin L, Whithead S, Dupuy B, Dougan G, Barrell B, Parkhill J. The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. Nat Genet. 2006;38(7):779–86. doi:10.1038/ng1830.

Saheb Kashaf *et al. BMC Systems Biology* (2017) 11:25

Page 13 of 13

51. Karasawa T, Ikoma S, Yamakawa K, Nakamura S. A defined growth medium for Clostridium difficile. Microbiology. 1995;141(2):371–5. doi:10.1099/13500872-141-2-371.

52. Janoir C, Denève C, Bouttier S, Barbut F, Hoys S, Caleechum L, Chapetón-Montes D, Pereira FC, Henriques AO, Collignon A, Monot M, Dupuy B. Adaptive strategies and pathogenesis of Clostridium difficile from In vivo transcriptomics. Infect Immun. 2013;81(10):3757–69. doi:10.1128/IAI.00515-13.

53. Kleessen S, Laitinen R, Fusari CM, Antonio C, Sulpice R, Fernie AR, Stitt M, Nikoloski Z. Metabolic efficiency underpins performance trade-offs in growth of Arabidopsis thaliana,. Nat Commun. 2014;5:3537. doi:10.1038/ncomms4537.

54. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10. doi:10.1093/nar/30.1.207.

55. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U, Brazma A. ArrayExpress update-simplifying data submissions. Nucleic Acids Res. 2015;43(D1):1113–6. doi:10.1093/nar/gku1057.

56. Gerber M, Walch C, Löffler B, Tischendorf K, Reischl U, Ackermann G. Effect of sub-MIC concentrations of metronidazole, vancomycin, clindamycin and linezolid on toxin gene transcription and production in Clostridium difficile. J Med Microbiol. 2008;57(6):776–83. doi:10.1099/jmm.0.47739-0.

57. Cattoir V, Ould-Hocine ZF, Legrand P. Antimicrobial susceptibility of Clostridium difficile clinical isolates collected from 2001 to 2007 in a French university hospital. Pathologie Biologie. 2008;56(7-8):407–11. doi:10.1016/j.patbio.2008.07.022.

58. Beran V, Chmelar D, Vobejdova J, Konigova A, Nemec J, Tvrdik J. Sensitivity to antibiotics of Clostridium difficile toxigenic nosocomial strains. Folia Microbiologica. 2014;59(3):209–15. doi:10.1007/s12223-013-0283-1.

59. Emerson JE, Stabler RA, Wren BW, Fairweather NF. Microarray analysis of the transcriptional responses of Clostridium difficile to environmental and antibiotic stress. J Med Microbiol. 2008;57(6):757–64. doi:10.1099/jmm.0.47657-0.

60. Jozefczuk S, Klie S, Catchpole G, Szymanski J, Cuadros-Inostroza A, Steinhauser D, Selbig J, Willmitzer L. Metabolomic and transcriptomic stress response of Escherichia coli. Mol Syst Biol. 2010;6(1):364. doi:10.1038/msb.2010.18.

61. Dubois T, Dancer-Thibonnier M, Monot M, Hamiot A, Bouillaut L, Soutourina O, Martin-Verstraete I, Dupuy B. Control of Clostridium difficile physiopathology in response to cysteine availability. Infect Immun. 2016;84(8):2389–405. doi:10.1128/IAI.00121-16.

62. Juhas M, Eberl L, Church GM. Essential genes as antimicrobial targets and cornerstones of synthetic biology. Trends Biotechnol. 2012;30(11):601–7. doi:10.1016/j.tibtech.2012.08.002.

63. Dembek M, Barquist L, Boinett CJ, Cain AK, Mayho M, Lawley TD, Fairweather NF, Fagan RP. High-throughput analysis of gene essentiality and sporulation in Clostridium difficile. mBio. 2015;6(2). doi:10.1128/mBio.02383-14.

64. Tatusova T, Dicuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016;44(14):6614–24. doi:10.1093/nar/gkw569.