METHOD ARTICLE

REVISED Elucidating genomic gaps using phenotypic profiles [version 2; peer review: 1 approved, 1 approved with reservations]

Daniel A. Cuevas[1], Daniel Garza[2,6], Savannah E. Sanchez[3], Jason Rostron[3], Chris S. Henry[4], Veronika Vonstein[5], Ross A. Overbeek[5], Anca Segall[3], Forest Rohwer[3], Elizabeth A. Dinsdale[3], Robert A. Edwards[1-5]

[1]Computational Science Research Center, San Diego State University, San Diego, CA, 92182, USA
[2]Department of Computer Science, San Diego State University, San Diego, CA, 92182, USA
[3]Department of Biology, San Diego State University, San Diego, CA, 92182, USA
[4]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 60439, USA
[5]Fellowship for Interpretation of Genomes, Burr Ridge, IL, 60527, USA
[6]Environmental Microbiology Laboratory, Evandro Chagas Institute, Ananindeua-PA, Brazil

v2 **First published:** 04 Sep 2014, **3**:210
https://doi.org/10.12688/f1000research.5140.1
Latest published: 17 Oct 2016, **3**:210
https://doi.org/10.12688/f1000research.5140.2

Abstract

Advances in genomic sequencing provide the ability to model the metabolism of organisms from their genome annotation. The bioinformatics tools developed to deduce gene function through homology-based methods are dependent on public databases; thus, novel discoveries are not readily extrapolated from current analysis tools with a homology dependence. Multi-phenotype Assay Plates (MAPs) provide a high-throughput method to profile bacterial phenotypes by growing bacteria in various growth conditions, simultaneously. More robust and accurate computational models can be constructed by coupling MAPs with current genomic annotation methods. *PMAnalyzer* is an online tool that analyzes bacterial growth curves from the MAP system which are then used to optimize metabolic models during *in silico* growth simulations. Using *Citrobacter sedlakii* as a prototype, the Rapid Annotation using Subsystem Technology (RAST) tool produced a model consisting of 1,367 enzymatic reactions. After the optimization, 44 reactions were added to, or modified within, the model. The model correctly predicted the outcome on 93% of growth experiments.

Keywords

high-throughput model reconciliation

Open Peer Review

Approval Status ? ✓

	1	2
version 2 (revision) 17 Oct 2016		✓ view
version 1 04 Sep 2014	? view	? view

1. **Matthew A. Oberhardt**, Tel Aviv University, Tel Aviv, Israel

2. **Aaron Best**, Hope College, Holland, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Daniel A. Cuevas (dcuevas08@gmail.com), Robert A. Edwards (raedwards@gmail.com)

Competing interests: No competing interests were disclosed.

Grant information: This work is partially supported by NSF grants CNS-1305112 and MCB-1330800 to Edwards, DUE-132809 to Dinsdale, DEB-1046413 to Rohwer, and by a STEM scholarship award funded by NSF grant DUE-1259951 to Cuevas.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Cuevas DA, Garza D, Sanchez SE *et al.* **Elucidating genomic gaps using phenotypic profiles [version 2; peer review: 1 approved, 1 approved with reservations]** F1000Research 2016, **3**:210 https://doi.org/10.12688/f1000research.5140.2

First published: 04 Sep 2014, **3**:210 https://doi.org/10.12688/f1000research.5140.1

REVISED **Amendments from Version 1**

Version 2 contains changes in the text and figures that clarify the ambiguities recognized by the referees. Figure 1, Figure 2, and Figure 3 have been revised. Supplementary Figure 4 has been added based on suggestions by the referees. We have also made available the metabolic models in SBML format. We thank the referees for their time and comments.

See referee reports

Introduction

Recent advancements in genomic sequencing provide high quality, deep-coverage DNA sequences for tens of thousands of bacterial genomes. To manage this breadth of data, online tools such as RAST (http://rast.nmpdr.org/)[1] leverage the SEED database[2] by using homology and genomic context to determine gene functions encoded in the DNA sequences. This automated annotation service additionally generates a raw metabolic reconstruction of the genome for use in *in silico* experiments. Genome-scale metabolism analyses use these reconstructions as input into data environments such as KBase, the Department of Energy Systems Knowledgebase (http://kbase.us). Several hypotheses can be tested simultaneously, e.g., protein function identification, biological behavior simulations, and metabolic network comparisons[3].

Metabolic models are defined by the chemical reactions that characterize the vast metabolic network of an organism. Flux-balance analysis (FBA) uses these chemical reactions to provide understanding of the physiological capacity of the cell[4]. Mathematically, the stoichiometry of metabolic networks is represented by a two-dimensional numerical matrix, in which the values are the stoichiometric coefficients of the reactants and products. Each row and column in the matrix is associated with a metabolite and a metabolic reaction, respectively. For one stoichiometric reaction, the products of the reaction are given positive integers, the reactants are given negative integers, and non-associated metabolites are given zeros. Through a constraint-based approach, the FBA algorithm uses linear programming techniques to solve this system of stoichiometric coefficients, optimizing for biomass production or another objective function[4–6].

The amount of published metabolic reconstructions for prokaryotic and eukaryotic organisms has increased over the past decade[3,7]. Through the increased use of next-generation sequencing and automated annotation software, metabolic models for new organisms are arising and older models are continuously being reconciled. However, a drawback to RAST and other gene annotation algorithms is the dependency on previous functional annotations. The breadth and quality of annotated functions vary among and across bacterial species, which is not accumulating as quickly as new sequences. Automatic generation of metabolic models is limited by our knowledge of cellular metabolism and biochemistry. In addition, an existing problem with gene databases is the inconsistent nomenclature used to name and define the function of a gene. Separate databases hold slightly different annotations for the same gene, which propagates into downstream tools, leading to a loss of information in analyses as mis-annotations cause reactions to be missing in models built during the initial reconstruction. To bridge the gap between quality genome annotations and accurate metabolic models novel methods are needed to supplement the reconciliation process.

Multi-phenotype Assay Plates (MAPs) provide a system to quantitatively monitor microbial growth while qualitatively deducing the metabolic capabilities of a microbe across a range of conditions. The MAPs technology uses optical density to measure biomass production by substrate utilization of a clonal microbial population. MAPs and similar technologies, such as Biolog's Phenotype MicroArrays, have been an advantageous tool in past phenotypic studies[3,8–20,29]. Although the Biolog system similarly measures substrate utilization, it does so as a response of cell respiration (i.e., reduction of a tetrazolium dye) and not specifically biomass production. Using MAPs, we can measure bacterial growth in defined conditions to aid in the validation of microbial genome annotation software.

In this study, a metabolic model of *Citrobacter sedlakii* is built using a workflow combining experimental data and computational analysis (Figure 1). The genome of *C. sedlakii* was sequenced, annotated, and a metabolic reconstruction was subsequently generated using RAST and the KBase platform. Growth of *C. sedlakii* was measured in 96 different growth conditions and the resulting data was introduced into a novel computational pipeline, PMAnalyzer (https://vdm.sdsu.edu/pmanalyzer). The PMAnalyzer automatically parameterizes raw growth data and fits a logistic model of bacterial growth for each experimental condition. Observed phenotypes from the MAP experiments were used to ground truth the genome-scale metabolic model by running FBA simulations on the KBase platform, which identified disparities in the metabolic reconstruction. Disparities were observed to have either been missed due to RAST mis-annotations or sequencing. Here, we introduce a high-throughput workflow to obtain large-scale metabolic reconstructions and reconciliations with observed growth phenotypes.

Methods

Acquisition of *Citrobacter sedlakii* and MAP (Multi-phenotype Assay Plate) preparation

The *C. sedlakii* isolate (ATCC 51115, CDC 4696-86) was provided by Dr. Marlene DeMers in the Department of Biology at San Diego State University. A glycerol frozen stock of the sample was plated on trypticase soy agar (Becton, Dickenson and Company) and incubated at 37°C for 24 hrs. A single colony was inoculated into 3 mL of trypticase soy broth (Becton, Dickenson and Company) and incubated, with shaking, at 37°C for 24 hrs. 500 µL of the overnight culture was mixed with 500 µL of 30% (weight/volume) filter sterilized glycerol and transferred to a cryogenic vial (Fisher Scientific) for storage at -80°C.

C. sedlakii was grown overnight at 37°C for 24 hrs on 50% Luria-Bertani (LB) agar (Fisher Scientific) from a frozen glycerol stock. Three independent colonies, biological triplicates, were inoculated into 3 mL of a modified 3-morpholinopropane-1-sulfonate (MOPS) broth[21] (1X MOPS (40 mM MOPS + 10 mM Tricine), 0.4% glycerol, 9.5 mM NH_4Cl, 0.25 mM $NaSO_4$, 1.0 mM $MgSO_4$, 1.32 mM K_2HPO_4, 10 mM KCl, 0.5 µM $CaCl_2$, 5 mM NaCl, and 6 µM $FeCl_3$) and incubated for 24 hrs at 37°C with agitation (250 rpm). Overnight cultures were centrifuged using

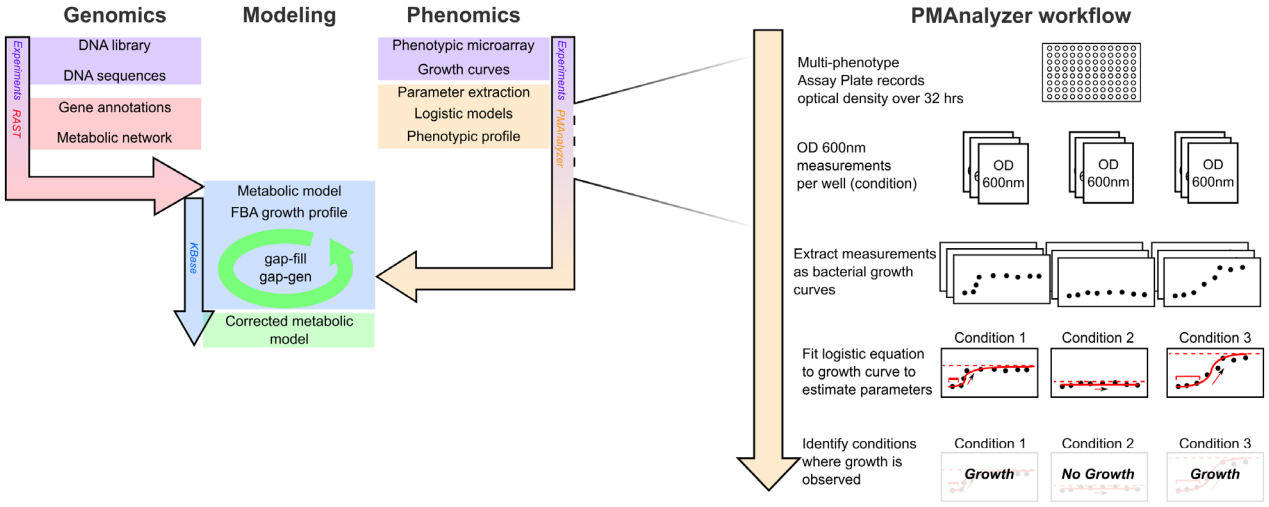


Figure 1. Combined flowchart. Analysis pipeline for generating and reconciling metabolic models. Initial metabolic models are built and reconciled in KBase from RAST genome annotations. Phenotypic profiles from the MAPs technology and the PMAnalyzer are incorporated into the KBase FBA-reconciliation loop to optimize the model. On the right are details of the PMAnalyzer workflow portraying the production of growth profiles from the MAPs.

an Eppendorf Centrifuge 5418R at maximum speed (14,000 rpm) to pellet cells and washed with 500 µL of 10 mM Tris/10 mM $MgSO_4$ buffer, twice. Cells were re-suspended in 1 mL of 10 mM Tris/10 mM $MgSO_4$ buffer and optical density at 600 nm (OD_{600}) was measured using a Beckman Coulter DU 640 spectrophotometer. All suspensions were concentrated to achieve a final optical density of approximately OD_{600} = 0.1 after a fifteen fold dilution.

10 µL of concentrated cells was transferred into each well of a sterile 96 well, micro-titer plate (Grenier Biosciences), which contained 60 µL of sterile water, 50 µL of 3X MOPS basal media, and 30 µL of 5X substrate (Supplementary Figure 1). Each plate was sealed with PCR grade plate film (Sigma SealPlate® film) with gas exchange still possible, and incubated on a Molecular Devices Analyst GT multi-plate plate reader (Molecular Devices, LLC.). Plate reader was programmed to incubate MAPs at 37°C and measure OD_{600} every 30 min with shaking before each read, for a total of 32 hrs. Absorbance data was saved, extracted as a text file, and uploaded to the project website for data storage (http://vdm.sdsu.edu/).

MOPS basal media is derived from the culture media provided by Neidhardt *et al.*[21], and contains 1X MOPS (40 mM MOPS + 10 mM Tricine), 0.4% glycerol*, 9.5 mM NH_4Cl*, 0.25 mM $NaSO_4$*, 1.0 mM $MgSO_4$*, 1.32 mM K_2HPO_4*, 10 mM KCl, 0.5 µM $CaCl_2$, 5 mM NaCl, 6 µM $FeCl_3$. Media was prepared with sterile Milli-Q water (Milli-Q Integral Water Purification Systems, EMD Millipore) and subsequently filter sterilized using 0.22 µm Sterivex filter unit (Millipore, Inc). (*These compounds are not included depending on the basal media. For example, 0.4% glycerol is not used in the carbon basal media, while 1.0 mM $MgSO_4$ is replaced by 1.0 mM $MgCl_2$ in the sulfur basal media).

Nutrient substrates were prepared by dissolving 1.25% (w/v) of the solid compound in sterile Milli-Q water and filter sterilized with a 0.22 µm Sterivex filter unit (Millipore, Inc). Substrate stocks were stored at 5X concentrations at room temperature in sterile conical tubes. Supplementary Figure 1 contains a detailed mapping of the substrates used in the MAPs.

Sequencing and metabolic reconstruction of *C. sedlakii*

As part of a DNA sequencing class at San Diego State University[26] the *C. sedlakii* 119 genome was sequenced using 454 pyrosequencing with the GS Junior platform and assembled with Newbler version 2.7. RAST (http://rast.nmpdr.org/) was used for subsystem annotations and metabolic reconstructions[1]. Annotations were imported into the KBase environment where the metabolic model was viewed, manipulated, and used in flux-balance analysis (FBA) simulations. FBA was used to determine if the model bacteria is successful in ascertaining growth in specific conditions equivalent to the MAPs. The KBase command *kbfba-importfbamodel* was used to import the annotations into the *Citrobacter_sedlakii_119* workspace and was named *C.sedlakii_nogapfill*. The model is also available in the supplementary files in SBML format in Dataset 1. The *Citrobacter_sedkakii_119* workspace and its objects are freely accessible to anyone. Using the KBase command *kbfba-gapfill*, initial gap-filling was performed on the model while specifying Luria-Bertani (LB) as the growth condition (the default ArgonneLB-Media formulation was used). The reconciled model was named *C.sedlakii_ArgonneLB_gapfill* and is provided in the workspace. This gap-filled model is available in the supplementary material in SBML files in Dataset 1. The LB gap-filled model created a representative model that fulfills the general requirements needed to utilize a rich media source for growth.

PMAnalyzer pipeline

The high-throughput analysis pipeline described below, and in Figure 1, was executed in a Linux command-line environment and was developed using several programming languages, including bash, Perl version 5.16 (http://www.perl.org/), and Python version 3.4.1 (http://www.python.org/). Perl scripts were written to parse and format the MAP raw data files into the tab-delimited intermediate files. The primary analysis script (Python) used these intermediate files for modeling the growth curves. For ease of execution, a single bash program was created as a wrapper script that executes the parsing and analysis scripts as a cohesive, automated pipeline. Command-line arguments or a configuration file was used for user input and settings. All scripts are freely accessible from a Git repository at https://github.com/dacuevas/PMAnalyzer. The online implementation can be found at https://vdm.sdsu.edu/pmanalyzer.

Fitting a logistic model to absorbance data

Phenotypic responses were recorded by measuring the optical density at 600 nm (OD_{600}) over time, which quantitatively represents the bacterial biomass concentration at each time point. The OD_{600} values are plotted to form the sigmoidal shape characteristic of bacterial growth curves. This characteristic curve, highlighted by Monod[22] and modeled by Zwietering *et al.*[23], consists of three phases: lag, exponential, and stationary phases. Zwietering *et al.*[23] interprets these phases as parameters required to model growth, using his logistic equation

$$\hat{y} = y_0 + \frac{A - y_0}{1 + \exp\left[\frac{\mu}{A}(\lambda - t_i) + 2\right]} \tag{1}$$

where y_0 (OD_{600}) is the starting optical density, λ (hr) is the lag phase, μ (OD_{600}•hr^{-1}) is the maximum growth rate during the exponential phase, A (OD_{600}) is the asymptote of the growth curve representing the carrying capacity of the population, and t (hr) is time. Supplementary Figure 2 provides a visual representation of a classical growth curve.

To parameterize the growth curves median values of the replicates were used. Python's NumPy module version 1.8.1 and SciPy module version 0.14.0[24] provides several functions for optimizing nonlinear, multivariate functions. In this case, the minimize function was used in order to denote bounds and constraints on each parameter. The default Broyden, Fletcher, Goldfarb, and Shanno (BFGS) algorithm[25] was used to minimize the sum of squared error between the logistic model from (1) and the raw data. As input, the algorithm requires estimations for each growth curve phase. The estimation for the asymptote was defined as the largest OD_{600} reading from three consecutive time points (2) and the maximum growth rate was defined as the largest change in OD_{600} over a 1.5 hrs window (3). The estimated lag time was set at 0.5 hr.

$$A = \max[\text{avg}(y_i, y_{i+1}, y_{i+2})] \tag{2}$$

$$\mu = \max\left[\frac{\log(y_{i+3}) - \log(y_i)}{t_{i+3} - t_i}\right] \tag{3}$$

The result from (2) was also used as the upper bound for the minimization function. Lag time and maximum growth rate were not given an upper bound. Lower bounds for the asymptote, maximum growth rate, and lag time were 0.01, 0, and 0, respectively.

Determining growth classifications

Each well in the MAPs has a varying level of growth, including different lag times, maximum growth rates, and asymptotes. A single value that represents the overall level of bacterial growth per well was generated by adapting the logistic model with the asymptote:

$$growth = \frac{n}{\sum_i^n \frac{1}{x_i}}, \quad where\, x_i = \hat{y}_i + A \tag{4}$$

Here, $\hat{y}_i$ is the value from (1) at time i, and n is the number of data values used, which is the number of OD_{600} measurements recorded during the experimental run. For the *C. sedlakii* MAP, n equals 64. The asymptote factor A, rather than the maximum growth rate, contributes to defining growth (i.e., growth levels from wells that achieve a higher biomass yield separate from those growth levels of wells that exhibit less growth, Supplementary Figure 3). In certain instances, growth curve models were fitted with a high maximum growth rate but did not display growth (e.g., potassium sorbate, L-valine, L-lysine, L-leucine, D-aspartic acid, and L-isoleucine). Ultimately, (4) was implemented to distill each growth curve into a single boolean variable of *growth* (≥ 0.5) or *no growth* (< 0.5).

Model reconciliation using the MAPs

The *kbfba-simpheno* function in KBase executes multiple flux-balance analysis (FBA) processes in parallel. To perform this, a text file listing information on which media condition to use as the input media in each process is required. Information regarding the PMAnalyzer result, i.e., growth or no growth, for each media condition tested on the MAPs are also listed in the text file. Digital representations of rich LB media and 90 different media compositions used in the MAPs were generated as media data objects in KBase. Each media object represents a specific condition used in the MAPs. *kbfba-simpheno* performs a separate FBA on each media object and compares the result to the MAPs result listed in the information text file. KBase FBA results are labeled as: *Correct Positive* assertions (FBA and MAPs both display growth), *Correct Negative* assertions (FBA and MAPs both display no growth), *False Positive* assertions (FBA asserts growth, MAPs display no growth), and *False Negative* assertions (FBA asserts no growth, MAPs display growth). Gap-fillings were attempted for conditions associated with false negative assertions and gap-generations were attempted for instances of false positive assertions. As stated previously, FBA and reconciliation was performed on the LB condition first in order to identify and integrate missing reactions required for growth on general, rich media. Thereafter, using the base model capable of asserting growth on LB, FBA and reconciliation was performed on the minimal media conditions with false negative assertions to target missing reactions in specific metabolic pathways. The minimal set of reactions determined by gap-filling was integrated into the model using the KBase function *kbfba-integratesolution*. To verify that the integration of new reactions produces additional correct positives and correct negatives the multi-FBA simulation was re-executed. Subsequently after gap-filling and multi-FBA simulation, some growth conditions resulted as false positives. Conditions where the model correctly asserted no growth prior to gap-filling had changed to incorrectly asserting growth. The KBase function

fba-gapgen was executed on those conditions to identify a set of reactions to alter or remove from the model that will resolve the false positive into a correct negative result. This algorithm attempts this without altering the outcome of a correct positive result condition, thus ensuring critical reactions for the model to grow are not removed.

Genomic analysis of gap-filled reactions

The gap-filling process provides biochemical processes that are missing from the initial model, and understanding why these reactions were missing during the initial reconstruction is important to investigate. The genomic databases available are numerous and nomenclature can differ between these resources. Reaction names for functions of genes may vary between databases, thus, resulting in a disconnect between gene annotations and the biochemical reactions those genes are involved in. When identifying functions for metabolic reconstructions, this unclear nomenclature prevents some reactions to appear in the initial metabolic model, thus producing a "mis-annotation". To correct for this, following gap-filling, all missing reactions (excluding transporters and newly-modified bidirectional reactions) were cross-checked with the SEED database to find similarly named reactions. The new list consisted of a mapping of gap-filled reaction names to possible alternative names. This list of reactions was then referenced back to the *C. sedlakii* RAST annotations in order to determine if the reaction was identified by RAST as the alternative name but not included in the metabolic model. A successful match between an alternative name (from the cross-check list of gap-filled reactions) and a name from the original RAST annotations would mean the KBase system failed to include a reaction whose enzyme was identified during annotation, and therefore, should have been included in the initial metabolic model. When the search similar nomenclature did not resolve, a search for gap-filled reactions in closely related organisms was performed; i.e., *Citrobacter koseri* and *E. coli* K12. This consisted of Protein BLAST (blastp) searches of the reactions' sequences from the SEED database against *C. koseri* and *E. coli*. Gap-filled reactions that are present in *C. sedlakii*'s closely related genomes were included in the *C. sedlakii* model with high confidence since closely related taxonomic groups contain common genetic material and function.

The genes encoding the missing reactions may be present in low quality DNA sequences or low coverage genomic regions. Following sequence assembly, these sequences are not present within the contigs, preventing RAST from annotating the proposed function. However, neighboring genes or protein complexes may be present and annotated, suggesting that the gene in question is there but was poorly sequenced. From within related organisms, the protein sequences of these complexes were identified and searched against the *C. sedlakii* genome. Finding matches for neighboring genes increases the confidence of including the reaction into the model. This method is also applicable to those genes that were not sequenced or that fall between assembled contigs.

Results

C. sedlakii 119 was assembled into 320 contigs containing 4,604,104 nucleotides with an N_{50} of 28,039 bp. RAST annotated the genome as containing 4,035 protein encoding genes and 76 tRNA genes over 537 different subsystems (Figure 2). Hypothetical proteins consisted of 817 (~20%) of the protein coding sequences. Membrane transport features constituted 150 out of 3,031 subsystem features. RAST listed *E. coli* as a close neighbor and *Citrobacter koseri* as the closest *Citrobacter* neighbor. A BLASTn alignment (expected value of 10^{-4})[27] against the *C. koseri* ATCC BAA-895 genome (NC_009792.1) resulted in a whole genome coverage of 69.8% (69.6% when including plasmids (NC_009793.1 and NC_009794.1). Figure 2 also displays subsystems for the *C. koseri* genome and the *Escherichia coli* K12 genomes, two close phylogenetic neighbors of *C. sedlakii* whose genomes are publicly available in the SEED database.

Growth on MAPs of each biological triplicate was followed for 32 hrs on separate days. After completion, the OD_{600} readings were processed through the PMAnalyzer pipeline (including data parsing, curve-fitting, and growth level analysis) in 8 seconds. Using a 0.5 growth level (4) cutoff and manual inspection of curves falling under the cutoff, 48 out of the 90 growth conditions – 35 carbon-based media and 13 nitrogen-based media – exhibited growth (Figure 3 and Table 1).

The initial metabolic model generated in KBase (*C.sedlakii_nogapfill*) from RAST annotations contained 1,367 reactions and 1,277 substrates. An FBA simulation using this model and specifying LB as the media source resulted in no growth, however only eight reactions were added to simulate growth. These eight were back referenced to the RAST annotations to check for mis-annotations or functional annotations not included in the model. Two gap-filled reactions (dimethylallyl-diphosphate:isopentenyl-diphosphate [EC 2.5.1.1] and geranyl-diphosphate synthase [EC 2.5.1.1]) were mis-annotated by RAST as they both shared a function with another annotated reaction (geranyltranstransferase (farnesyldiphosphate synthase) [EC 2.5.1.10]). Undecaprenyl pyrophosphate synthetase [EC 2.5.1.31] and quinolinate synthase [EC 2.5.1.72] were also annotated and shared a function with the gap-filled reactions undecaprenyl diphosphate synthase [EC 2.5.1.31] and quinolinate synthetase [EC 2.5.1.72], respectively. The fifth reaction (1-deoxy-D-xylulose-5-phosphate pyruvate-lyase (carboxylating) [EC 2.2.1.7]) was not annotated but a comparison between the *C. koseri* and *E. coli* K12 genomes revealed neighboring homologs in *C. sedlakii* genome flanking a gap where the gene should be. Therefore, it is likely that RAST missed this gene during gene calling due to a frameshift sequencing error. Homologs of the proteins for three reactions (ATP:dTMP phosphotransferase [EC 2.7.4.9]; glutamine amidotransferase [EC 2.4.2.-]; riboflavin transport in/out via proton symport) were not found in *C. sedlakii* but were identified in the *C. koseri* genome. In addition to the previous eight reactions, three additional reactions (meso-2,6-diaminoheptanedioate carboxy-lyase [EC 4.1.1.20]; NADH:guanosine-5'-phosphate oxidoreductase (deaminating) [EC 1.7.1.7]; prephenate:NADP+ oxidoreductase (decarboxylating) [EC 1.3.1.13]) were altered from being uni-directional to bi-directional. The base model (*C.sedlakii_ArgonneLB_gapfill*), with the ability to grow on rich media, contained 1,279 (+2) substrates and 1,375 (+8) reactions.

Using the base model, the 90 well simulation resulted in no growth for all 90 growth conditions (53.3% accuracy: 48 false negatives

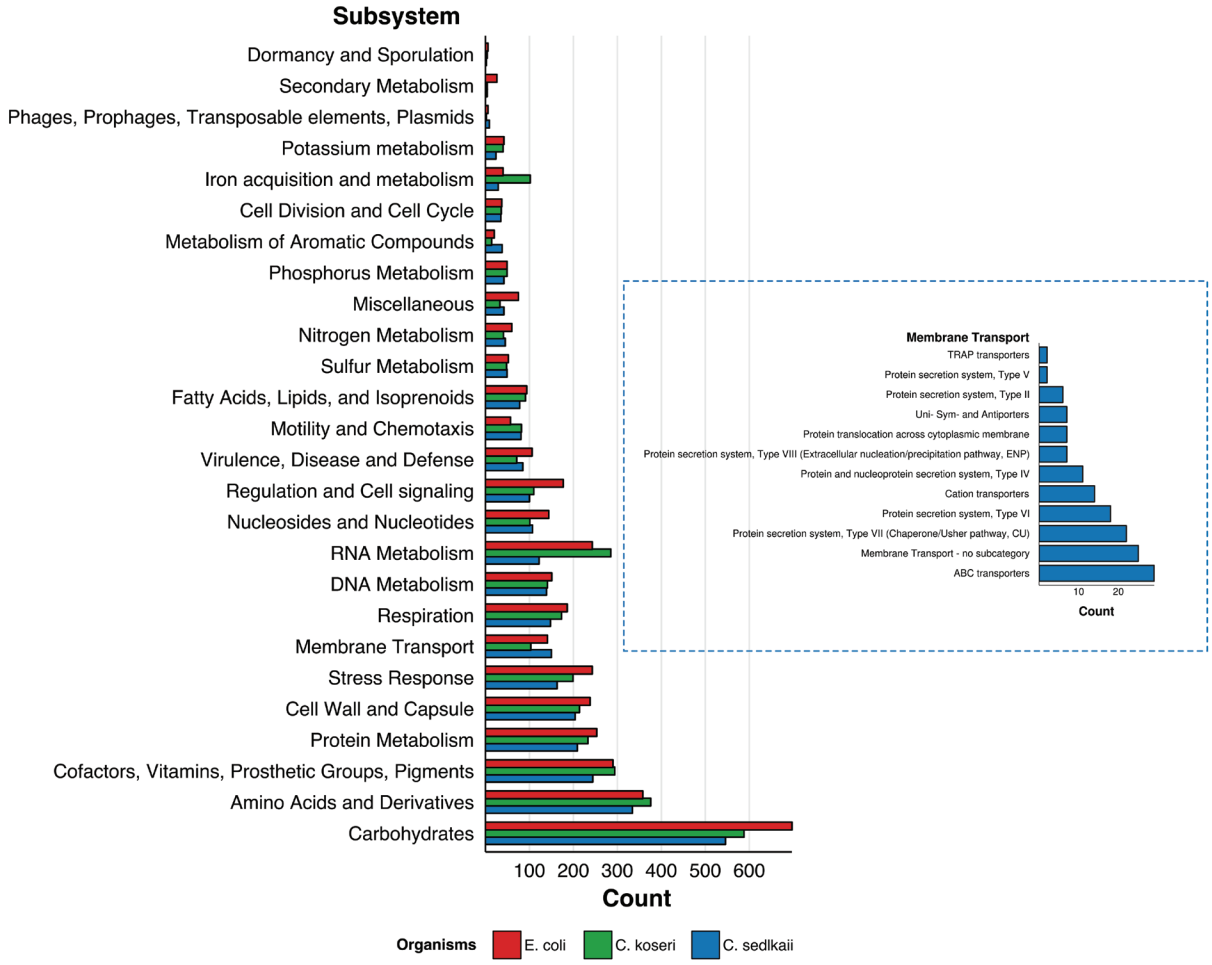


Figure 2. RAST subsystems. Number of subsystems annotated by RAST for each subsystem group for *Citrobacter sedlakii* and two phylogenetic neighbors *Citrobacter koseri* and *Escherichia coli*. Membrane transport subsystems identified in *C. sedlakii* are highlighted in separate plot.

(FN)) (Table 1). Subsequently, gap-filling was performed on those 48 conditions resulting in the reconciled model (*C.sedlakii_MOPS_simpheno*). The total number of substrates increased to 1,301 (+22 from the base model; +24 total), the total number of reactions increased to 1,399 (+24 from the base model; +32 total), and nine reactions were modified to be bi-directional. Transport reactions made up 11 of the 24 (46%) minimal media gap-filled reactions. When performing gap-filling on multiple conditions, KBase produced a separate solution for each condition. This results in reactions appearing in multiple solutions. To find the missing set of essential reactions, the gap-fill results were parsed to locate the minimum set of reactions present in the majority of the solutions. All new and modified reactions added to the model are listed in Table 2. Although Table 2 presents the gap-filled reactions for only 13 FN conditions, the addition of those gap-filled reactions correctly amended the other 35 FN conditions not shown in the table.

Mis-annotation checks and cross-referencing with the *C. koseri* genome were performed for the 12 gap-filled, non-transport, newly added reactions. In four cases (phosphoribosyl-ATP pyrophosphohydrolase [EC 3.6.1.31]; 1-(5-phospho-D-ribosyl)-AMP 1,6-hydrolase [EC 3.5.4.19]; D-mannose-6-phosphate ketol-isomerase [EC 5.3.1.8]; D-lactate dehydrogenase [EC 1.1.2.5]) the DNA sequences were missing but their surrounding genes were sequenced and annotated, indicative of poorly sequenced genes or genes located between contigs. Two reactions (chorismate pyruvatemutase [EC 5.4.99.5]; pyrimidine phosphatase [EC 3.1.3.-]) were mis-annotated by RAST. Four reactions (ureidoglycolate amidohydrolase (decarboxylating) [EC 3.5.3.19]; allantoate amidinohydrolase [EC 3.5.3.4]; allantoin amidohydrolase [EC 3.5.2.5]; 5-oxoproline amidohydrolase (ATP-hydrolysing) [EC 3.5.2.9]) could not be identified in *C. koseri* and two reactions (D-Arabinose ketol-isomerase [EC 5.3.1.3]; myo-inositol:oxygen oxidoreductase

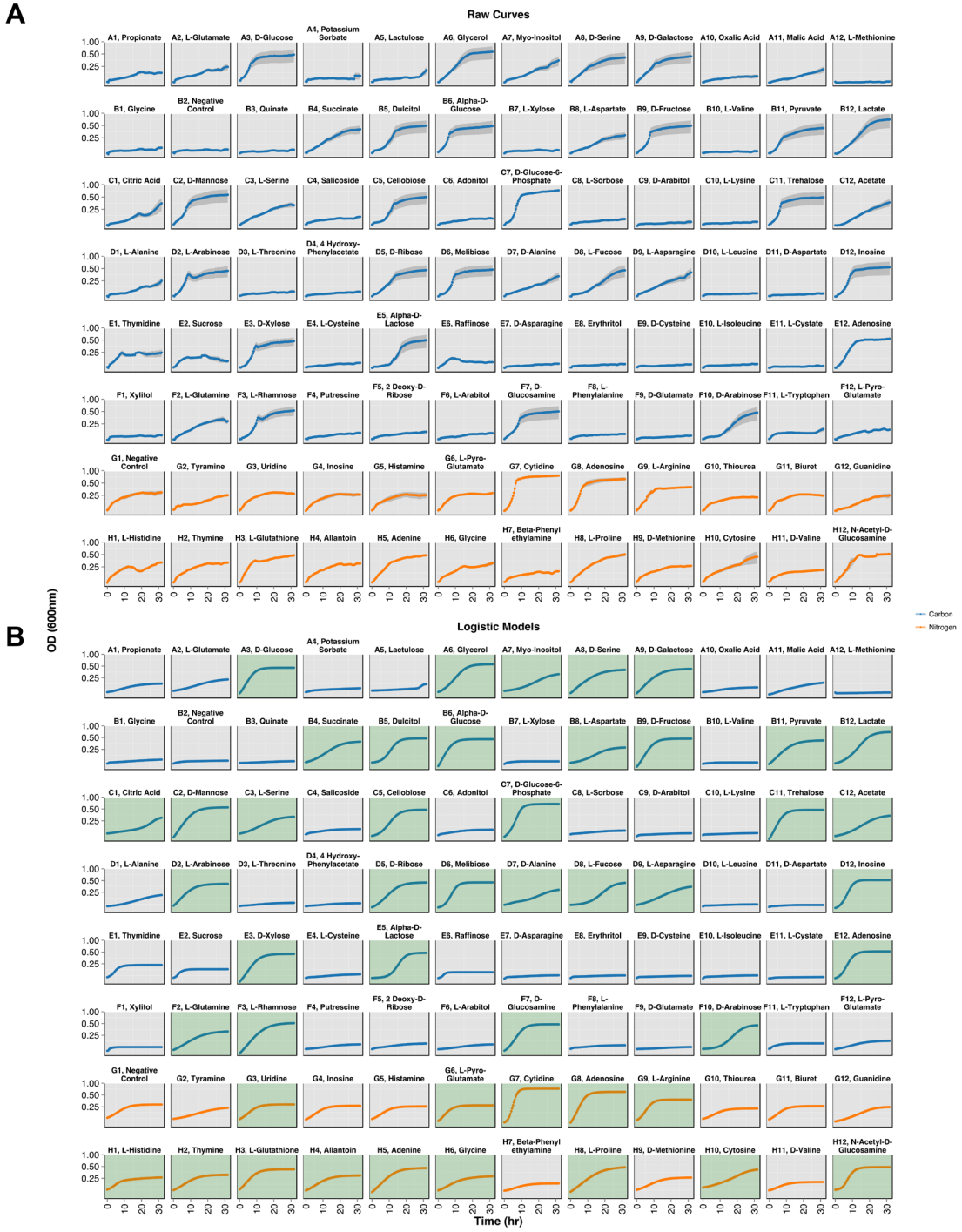


Figure 3. *Citrobacter sedlakii* growth curves. Growth curves generated from the multi-plate reader over 32 hr. The y-axis is displayed in a log 2 scale and substrate groups are distinguished by color: blue for carbon sources and red for nitrogen sources. (**A**) Standard error from the technical replicates are plotted as gray regions. (**B**) Logistic models of each growth condition. Green panels indicate conditions displaying growth.

Table 1. MAP and FBA results. *Citrobacter sedlakii* growth (G) and no growth (NG) phenotypes. Carbon substrates are denoted with (C); nitrogen substrates are denoted with (N). FBA results before gap-filling are shown in the Initial Model column. MAP-FBA comparison results are: correct positive (CP), correct negative (CN), false positive (FP), and false negative (FN).

Growth Condition	MAP	Initial Model	Multi Gap-Fill	Initial Result	Final Result
2 Deoxy-D-Ribose (C)	NG	NG	G	CN	FP
4 Hydroxy Phenyl Acetic Acid (C)	NG	NG	NG	CN	CN
Acetic Acid (C)	G	NG	G	FN	CP
Adenosine (C)	G	NG	G	FN	CP
Adonitol (C)	NG	NG	NG	CN	CN
Alpha-D-Glucose (C)	G	NG	G	FN	CP
Alpha-D-Melibiose (C)	G	NG	G	FN	CP
Citric Acid (C)	G	NG	G	FN	CP
D-Alanine (C)	G	NG	G	FN	CP
D-Arabinose (C)	G	NG	G	FN	CP
D-Arabitol (C)	NG	NG	NG	CN	CN
D-Asparagine (C)	NG	NG	NG	CN	CN
D-Aspartic Acid (C)	NG	NG	NG	CN	CN
D-Cellobiose (C)	G	NG	G	FN	CP
D-Cysteine (C)	NG	NG	NG	CN	CN
D-Fructose (C)	G	NG	G	FN	CP
D-Galactose (C)	G	NG	G	FN	CP
D-Glucosamine (C)	G	NG	G	FN	CP
D-Glucose (C)	G	NG	G	FN	CP
D-Glucose-6-PO4 (C)	G	NG	G	FN	CP
D-Glutamic Acid (C)	NG	NG	NG	CN	CN
D-Mannose (C)	G	NG	G	FN	CP
D-Raffinose (C)	NG	NG	NG	CN	CN
D-Ribose (C)	G	NG	G	FN	CP
D-Salicin (C)	NG	NG	NG	CN	CN
D-Serine (C)	G	NG	G	FN	CP
D-Trehalose (C)	G	NG	G	FN	CP
D-Xylose (C)	G	NG	G	FN	CP
Dulcitol (C)	G	NG	G	FN	CP
Erythritol (C)	NG	NG	NG	CN	CN
Glycerol (C)	G	NG	G	FN	CP
Glycine (C)	NG	NG	G	CN	FP
Inosine (C)	G	NG	G	FN	CP
L-Alanine (C)	G	NG	G	FN	CP
L-Arabinose (C)	G	NG	G	FN	CP
L-Arabitol (C)	NG	NG	NG	CN	CN
L-Asparagine (C)	G	NG	G	FN	CP
L-Aspartic Acid (C)	G	NG	G	FN	CP
L-Cysteic Acid (C)	NG	NG	NG	CN	CN
L-Cysteine (C)	NG	NG	NG	CN	CN
L-Fucose (C)	G	NG	G	FN	CP
L-Glutamic Acid (C)	G	NG	G	FN	CP
L-Glutamine (C)	G	NG	G	FN	CP
L-Isoleucine (C)	NG	NG	NG	CN	CN
L-Leucine (C)	NG	NG	NG	CN	CN
L-Lysine (C)	NG	NG	NG	CN	CN
L-Methionine (C)	NG	NG	NG	CN	CN
L-Phenyl-Alanine (C)	NG	NG	NG	CN	CN
L-Pyro-Glutamic Acid (C)	NG	NG	G	CN	FP
L-Rhamnose (C)	G	NG	G	FN	CP
L-Serine (C)	G	NG	G	FN	CP
L-Sorbose (C)	NG	NG	NG	CN	CN
L-Threonine (C)	NG	NG	G	CN	FP
L-Tryptophan (C)	NG	NG	NG	CN	CN
L-Valine (C)	NG	NG	NG	CN	CN
L-Xylose (C)	NG	NG	NG	CN	CN
Lactate (C)	G	NG	G	FN	CP
Lactulose (C)	NG	NG	NG	CN	CN
Malic Acid (C)	NG	NG	G	CN	FP
Myo Inositol (C)	G	NG	G	FN	CP
Oxalic Acid (C)	NG	NG	NG	CN	CN
Propionic Acid (C)	NG	NG	NG	CN	CN
Putrescine (C)	NG	NG	G	CN	FP
Quinic Acid (C)	NG	NG	NG	CN	CN
Sodium Pyruvate (C)	G	NG	G	FN	CP
Sodium Succinate (C)	G	NG	G	FN	CP
Sucrose (C)	NG	NG	NG	CN	CN
Thymidine (C)	G	NG	G	FN	CP
Xylitol (C)	NG	NG	NG	CN	CN
Adenine Sulfate (N)	G	NG	G	FN	CP
Adenosine 5' MonoPO4 (N)	G	NG	G	FN	CP
Allantoin (N)	G	NG	G	FN	CP
Beta Phenyl-Ethylamine (N)	NG	NG	NG	CN	CN
Biuret (N)	NG	NG	NG	CN	CN
Cytidine (N)	G	NG	G	FN	CP
Cytosine (N)	G	NG	G	FN	CP
D-Methionine (N)	NG	NG	NG	CN	CN
D-Valine (N)	NG	NG	NG	CN	CN
Glycine (N)	G	NG	G	FN	CP
Histamine (N)	NG	NG	NG	CN	CN
Inosine (N)	NG	NG	NG	CN	CN
L-Arginine (N)	G	NG	G	FN	CP
L-Histidine (N)	G	NG	G	FN	CP
L-Proline (N)	G	NG	G	FN	CP
L-Pyro Glutamic Acid (N)	G	NG	G	FN	CP
N-Acetyl-D-Glucosamine (N)	G	NG	G	FN	CP
Thiourea (N)	NG	NG	NG	CN	CN
Thymine (N)	G	NG	G	FN	CP
Tyramine (N)	NG	NG	NG	CN	CN
Uridine (N)	G	NG	G	FN	CP

Table 2. Gap-filled reactions. Each gap-filled reaction is listed under their respective media condition, along with the primary source of the compound in parenthesis, i.e. (C) denotes carbon and (N) denotes nitrogen. E.C. numbers were supplied by KBase when viewing gap-filling results. Reactions listed with an asterisk denote the 13 transport reactions added to the model. Reactions listed with a yes in the Reversible column denote those that were already present in the model but made bi-directional through gap-filling.

KBase Media	KBase Reaction ID	EC #	Reversible	Name
ArgonneLB	rxn00313	4.1.1.20	yes	meso-2,6-Diaminoheptanedioate carboxy-lyase
ArgonneLB	rxn00837	1.7.1.7;1.6.6.8	yes	NADH:guanosine-5'-phosphate oxidoreductase(deaminating)
ArgonneLB	rxn01213	2.5.1.1;2.5.1.29;2.5.1.10		Dimethylallyl-diphosphate:isopentenyl-diphosphate
ArgonneLB	rxn01269	1.3.1.43;1.3.1.13	yes	Prephenate:NADP+ oxidoreductase(decarboxylating)
ArgonneLB	rxn01466	2.5.1.1;2.5.1.29;2.5.1.10		Geranyl-diphosphate:isopentenyl-diphosphate geranyltrans-transferase
ArgonneLB	rxn01513	2.7.4.9;2.7.4.12		ATP:dTMP phosphotransferase
ArgonneLB	rxn02988			quinolinate synthase
ArgonneLB	rxn03909	2.2.1.7		1-Deoxy-D-xylulose-5-phosphate pyruvate-lyase (carboxylating)
ArgonneLB	rxn05144	2.4.2.-		Glutamine amidotransferase
ArgonneLB	rxn05645	*		riboflavin transport in/out via proton symport
ArgonneLB	rxn09345			Undecaprenyl diphosphate synthase
Acetate (C)	rxn01256	5.4.99.5		Chorismate pyruvatemutase
Acetate (C)	rxn02834	3.6.1.31		Phosphoribosyl-ATP pyrophosphohydrolase
Acetate (C)	rxn02835	3.5.4.19		1-(5-phospho-D-ribosyl)-AMP 1,6-hydrolase
Acetate (C)	rxn03031	2.3.1.117	yes	Succinyl-CoA:2,3,4,5-tetrahydropyridine-2,6-dicarboxylate
Acetate (C)	rxn05039	3.1.3.-		pyrimidine phosphatase
Acetate (C)	rxn10904	*		Acetate transport
D-Arabinose (C)	rxn01152	5.3.1.3		D-Arabinose ketol-isomerase
D-Arabinose (C)	rxn05499	*		D-arabinose reversible transport
D-Mannose (C)	rxn00559	5.3.1.8		D-Mannose-6-phosphate ketol-isomerase
L-Asparagine (C)	rxn05508	*		L-asparagine transport in via proton symport
L-Fucose (C)	rxn05691	*		L-Fucose transport in via proton symport
Lactate (C)	rxn11116	*		Lactate transport
Lactate (C)	rxn12893			D-lactate dehydrogenase [cytochrome] 1, mitochondrial
Myo-Inositol (C)	rxn00880	1.13.99.1		myo-Inositol:oxygen oxidoreductase
Myo-Inositol (C)	rxn05593	*		inositol transport in via proton symport
Pyruvate (C)	rxn05469	*		Pyruvate transport via proton symport
Adenine (N)	rxn00470	4.1.1.17	yes	L-Ornithine carboxy-lyase
Adenine (N)	rxn00509	1.2.1.16	yes	Succinate-semialdehyde:NADP+ oxidoreductase
Adenine (N)	rxn01851	1.2.1.3;1.2.1.19	yes	4-aminobutanal:NAD+ 1-oxidoreductase
Adenine (N)	rxn05491	*		adenine transport in via proton symport
Allantoin (N)	rxn00327	3.5.3.19		Ureidoglycolate amidohydrolase (decarboxylating)
Allantoin (N)	rxn01746	3.5.3.4		Allantoate amidinohydrolase
Allantoin (N)	rxn01748	3.5.2.5		Allantoin amidohydrolase
Allantoin (N)	rxn05682	*		allantoin transport in via proton symport
L-Histidine (N)	rxn12604	*		Urocanate transport via proton symport
L-Pyro-Glutamate (N)	rxn00186	3.5.2.9		5-Oxoproline amidohydrolase (ATP-hydrolysing)
L-Pyro-Glutamate (N)	rxn05694	*		L-Pyroglutamic Acid transport in via proton symport
Thymine (N)	rxn00710	4.1.1.23	yes	Orotidine-5'-phosphate carboxy-lyase
Thymine (N)	rxn01018	2.1.3.2	yes	Carbamoyl-phosphate:L-aspartate carbamoyltransferase
Thymine (N)	rxn01362	2.4.2.10	yes	Orotidine-5'-phosphate:pyrophosphate phosphoribosyltransferase
Thymine (N)	rxn01520	2.1.1.45	yes	5,10-Methylenetetrahydrofolate:dUMP C-methyltransferase
Thymine (N)	rxn08335	1.3.3.1	yes	dihydoorotic acid dehydrogenase (quinone8)
Thymine (N)	rxn09656	*		thymine reversible transport via proton antiport

[EC 1.13.99.1]) could not be identified in *C. koseri* nor in *E. coli*. Therefore, these six reactions were added with minimal evidence. From the six reactions, three reactions (including the transporter) were required for growth on allantoin. In fact, *Citrobacter* species have not been shown to be capable of utilizing allantoin as a sole source of nitrogen[28]. More in-depth experiments are needed to investigate *C. sedlakii*'s ability to grow in the allantoin-based growth condition.

The 90 simulations were executed again and resulted in growth on 54 conditions (48 TP, 6 FP) and no growth on 36 (36 TN, 0 FN), a final accuracy of 93.3%. The six false positive growth conditions were: 2 deoxy-D-ribose (carbon), glycine (carbon), L-pyroglutamic acid (carbon), L-threonine (carbon), malate (carbon), and putrescine (carbon) (Table 1). In an attempt to resolve the false positives, the KBase gap-generation algorithm was executed; however, it was unable to identify a set of reactions that could be safely altered or removed from the model without changing correct positive results into false negatives.

Phenotypic profiling data for elucidating genomic gaps

6 Data files

http://dx.doi.org/10.6084/m9.figshare.3969072.v1

Discussion

RAST annotations and gap-filled reactions

Back referencing the gap-filled reactions to the prior RAST annotations provides insight into KBase's ability to build the initial reconstruction from subsystem annotations. Several reactions identified by RAST were not included in the metabolic model. This is either due to RAST not being able to determine the function of the gene using homology, or the model is not able to correctly incorporate the function from the annotation. Gap-fill on LB media identified four reactions that were later categorized as RAST mis-annotations, whereas gap-fill on the MAPs media identified only two reactions later categorized as RAST mis-annotations. By surveying neighboring homologs of gap-filled reactions in closely related genomes, one reaction on LB gap-fill and four reactions on MAPs gap-fill were categorized as missing due to poor sequence quality.

As more bacterial genomes are processed through this pipeline, mis-annotations will occur at a lower rate. In the case of *C. sedlakii*, six reactions out of the 19 non-transport, newly added reactions were missed due to mis-annotations; five reactions possibly located in-between contigs; eight not bearing homology with related organisms. Feedback from mis-annotations corrects the ambiguities for future bacteria metabolic reconstructions as administrators of KBase, RAST, and the SEED database are informed of these findings. Reactions that are gap-filled and not identified as mis-annotations will be recorded for future studies. Patterns of missing reactions provide insight into why RAST is not able to identify the functions from an organism's sequences. Furthermore, closer investigation can answer if the gene encoding the protein for a particular species (or genus) does not share strong homology to its closest evolutionary neighbor.

Missing transporter proteins

Thirteen out of the 32 essential reactions (41%) added during gap-filling were transporters. For specific conditions, the only missing reaction that prevented growth during the simulation was observed to be transport proteins specific for the growth condition (see Table 2). Transporters are readily identified using homology-based searches, however, it is difficult to accurately identify which substrate(s) are actively transported using these techniques. A drawback of RAST is its dependence on sequence homology. Henry *et al.*[30] indicated that poorly annotated transporters are typically missing from preliminary reconstructions using the SEED database. During metabolic reconstruction in RAST, a minimal set of reactions are uniquely chosen through an optimization equation. The equation contains a penalty parameter that favors intracellular reactions over transporters during the auto-completion step of the model reconstruction, thus further preventing transporters from being included in the draft model.

FBA false positives

False positive (FP) results were introduced during the reconciliation process. A false positive was defined as an FBA resulting in biomass production in nutrient conditions where *C. sedlakii* is not able to produce biomass, i.e. the MAPs assert no growth. FPs are an indication of an under-constrained model that has resulted from the addition of reactions or from making reactions bi-drectional, both of which are enabling biomass production during FBA. In the case with the *C. sedlakii* model, FPs resulted after the gap-fill occurred. The 32 reactions added and nine reactions re-labeled as bidirectional have enabled growth in six growth conditions where the MAPs display no growth. KBase function *fba-gapgen* attempts to correct these issues. To perform this function, parameters pertaining to the growth condition where the FP occurs and a growth condition where a correct positive occurs are required, allowing the algorithm to correct, or remove, reactions from the model without altering the outcome of the correct positive growth condition. The KBase software was unable to determine any reactions in our model to remove with the *fba-gapgen* function, suggesting that the model requires manual curation and in-depth experimentation in order to resolve these six FP growth conditions listed in Table 1. Certain biochemical events not captured by the metabolic model may be occurring, such as the effect of transcription repression and other complex cellular behaviors. FBA alone is unable to account for such dynamic responses, although efforts have been made to extend genome-scale metabolic models with metabolic and gene expression data[34,35].

Models created using the discussed pipeline should be considered draft metabolic models. While the physiological experiments provide insight into an organism's metabolic capabilities for substrate utilization and the subsequent biomass formation, the biochemical properties involved are not directly assessed. The gap-filling algorithm attempts to determine the minimal set of reactions required for growth for a specific condition but these are not considered finite decisions. Gene knockout experiments, extensive literature mining, and manual curation are required to enhance the models[31]. However, these alternatives are neither high-throughput nor considered in this pipeline but are encouraged for further studies. Cross-referencing to closely related organisms can give insight into the validity of adding a reaction to the model[28].

Conclusion

The prevalence of complete and near-complete draft genomes is increasing as DNA sequencing becomes cheaper and more robust. Unique bacterial species are now being studied and their data is becoming more readily available for interpretation. We describe a process to combine DNA sequence data and phenotypic experiments to produce a programmatic metabolic model construct. Metabolic reconstructions are built using RAST genomic annotations by implementing PMAnalyzer, a high-throughput pipeline. Biochemical reactions not captured by homology-based algorithms are highlighted using flux-balance analysis and gap identification techniques hosted openly on the KBase platform. *Citrobacter sedlakii* was used as a model organism to describe, test, and critique the pipeline. FBA results using this model were shown to achieve a 93% prediction accuracy, an improvement from the initial model providing a 53% prediction accuracy.

The high-throughput pipeline presented combines physiological data with genomic information to result in more accurate metabolic models. Using experimentation to validate model prediction and improve model capabilities has been shown previously, however, these processes are not streamlined to be fast or robust. Our methodology implements speed and robustness at every level of the work-flow. Using a multi-plate spectrophotometer, hundreds of different growth conditions targeting several metabolic pathways results in an expansive phenotypic profile. The PMAnalyzer pipeline executes rapidly and is integrated into an automated web server where users obtain phenotypic results within minutes. Draft model construction and model reconciliation on KBase are completed quickly, providing gap-filled biochemical functions of a genome to use in subsequent functional analyses.

Data and software availability

Data

figshare: Phenotypic profiling data for elucidating genomic gaps. Doi: 10.6084/m9.figshare.3969072[32]

Software

Latest Software script source code: https://github.com/dacuevas/PMAnalyzer

Source code as at the time of publication: https://github.com/F1000Research/PMAnalyzer/releases/tag/V1.0

Archived source code as at the time of publication: http://dx.doi.org/10.5281/zenodo.11413[33]

Software License: GNU GPL v3.0

Author contributions

DAC and RAE conceived and designed the study. DAC wrote the paper, prepared figures and tables, developed and enhanced the PMAnalyzer code and pipeline, designed the experiments, performed experiments, and analyzed the data. DAC, DG, SES, and JR jointly discussed the design of the PMAnalyzer. SES designed the MAPs and performed the MAP experiments. ED provided DNA sequencer instrument, reagents, and sequences. JR, AS, and FR included helpful discussions with the MAPs. CSH contributed code to KBase. VV and RAO provided helpful discussions and suggestions with genome annotations.

Competing interests

No competing interests were disclosed.

Grant information

This work is partially supported by NSF grants CNS-1305112 and MCB-1330800 to Edwards, DUE-132809 to Dinsdale, DEB-1046413 to Rohwer, and by a STEM scholarship award funded by NSF grant DUE-1259951 to Cuevas.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

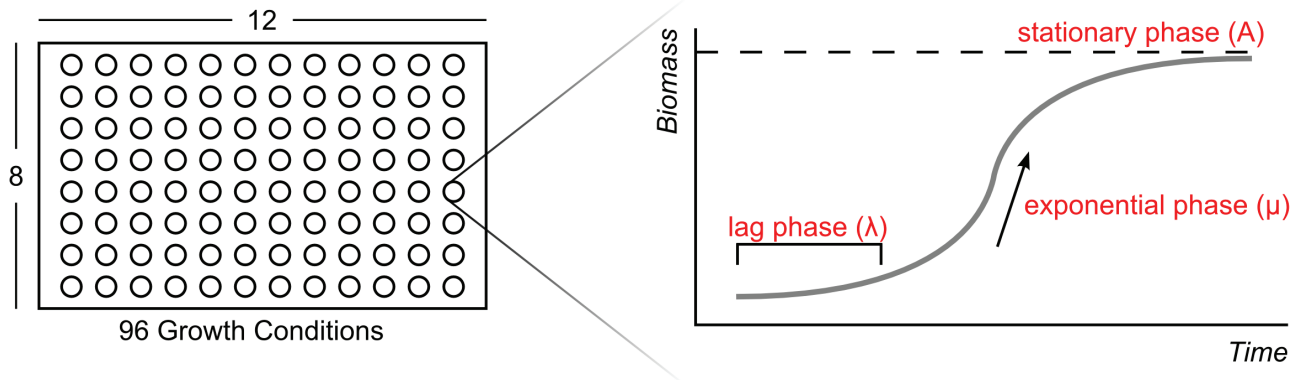
Acknowledgements

A special thank you to Barbara Bailey, Ben Felts, Jim Nulton, and Peter Salamon from the San Diego State University Bio Math group for their discussions and opinions with bacterial growth curve modeling.

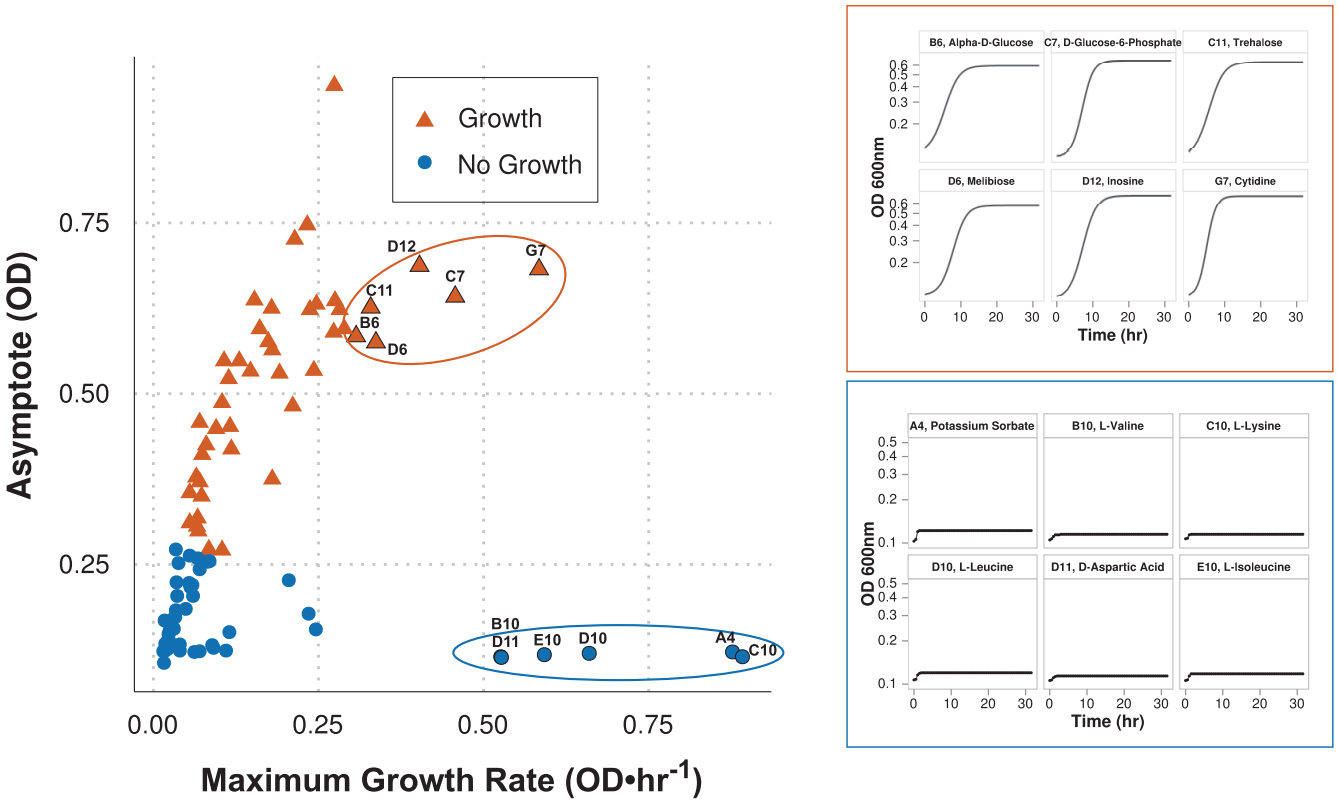
Supplementary material

	1	2	3	4	5	6	7	8	9	10	11	12
A	Propionate	L-glutamic acid	D-glucose	Potassium sorbate	Lactulose	Glycerol	Myo inositol	D-serine	D-galactose	Oxalic acid (0.2%)	Malic acid	L-methionine (0.2%)
B	Glycine	Water	Quinate	Succinate	Dulcitol	Alpha-d-glucose	L-xylose	L-aspartic acid (0.2%)	D-fructose	L-valine	Pyruvate	Lactate
C	Citric acid	D-mannose	L-serine	Salicoside	Cellobiose	Adonitol	D-glucose-6-$P0_4$	L-sorbose	D-arabitol	L-lysine	D-trehalose	Acetic acid
D	L-alanine	L-arabinose	L-threonine	4 hydroxy phenylacetate	D-ribose	Melibiose	D-alanine	L-fucose	L-asparagine (0.2%)	L-leucine	D-aspartic acid (0.2%)	Inosine
E	Thymidine	Sucrose	D-xylose	L-cysteine	Alpha-d-lactose	Raffinose	D-asparagine	Erythritol	D-cysteine	L-isoleucine (0.2%)	L-cysteic acid	Adenosine (0.2%)
F	Xylitol	L-glutamine	L-rhamnose	Putrescine	2 deoxy-d-ribose	L-arabitol	D-glucosamine	L-phenyl-alanine	D-glutamic acid (0.2%)	D-arabinose	L-tryptophan (0.2%)	L-pyro glutamic acid
G	Water	Tyramine (0.2%)	Uridine	Inosine	Histamine	L-pyro glutamic acid	Cytidine	Adenosine (0.2%)	L-arginine (0.2%)	Thiourea	Biuret (0.2%)	Guanidine
H	L-histidine	Thymine (0.2%)	L-glutathione (0.2%)	Allantoin (0.2%)	Adenine (0.2%)	Glycine	Beta phenyl-ethylamine	L-proline (0.2%)	D-methionine (0.2%)	Cytosine (0.2%)	D-valine (0.2%)	N-acetyl-d-glucosamine (0.2%)

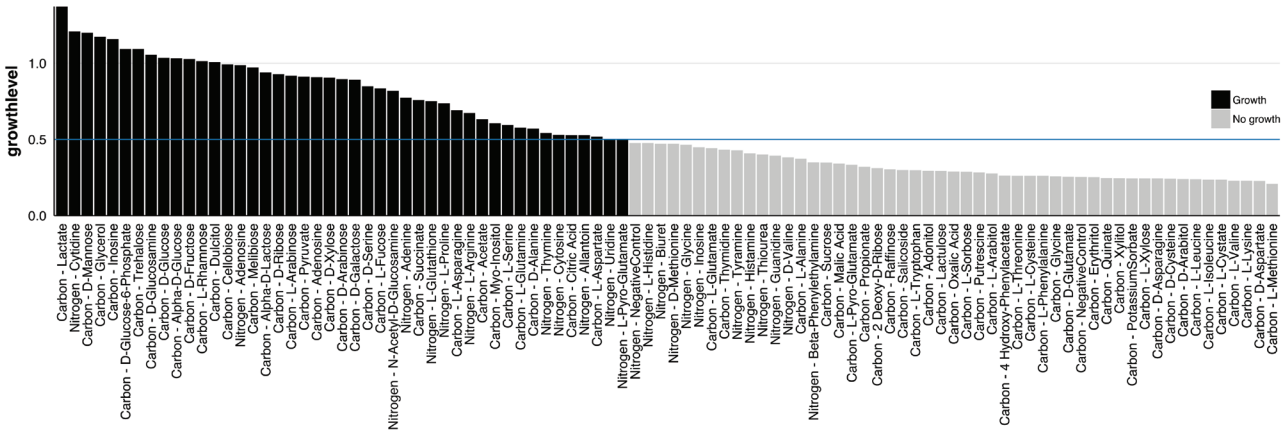
Supplemental Figure 1. Multi-phenotype Assay Plate. Each well contains 60 µL of sterile water, 50 µL of 3X MOPS basal media, and 30 µL of 5X substrate. Unless noted, a concentration of 0.25% is used for each substrate. Carbon substrates are denoted by blue squares in rows **A**–**F**; nitrogen substrates are denoted by red squares in rows **G** and **H**. Water denotes wells using water instead of a carbon (or nitrogen) substrate.



Supplemental Figure 2. Phenotypic assay and growth curve. A bacterial growth curve can be parameterized into three phases: lag phase, exponential phase, and stationary phase. Parameters correspond to the logistic equation described in the text (Equation 1).



Supplemental Figure 3. Growth vs. no growth. Classification of growth conditions are based on the asymptote adjusted logistic model calculation in Equation (4). Red triangles and blue circles indicate conditions asserting growth (≥ 0.5) and no growth (< 0.5), respectively. Classification is weighted by the final biomass yield (asymptote) rather than growth rate. Maximum growth rate can be misleading; e.g., growth curves highlighted in the blue box were modeled using high growth rates but do not assert growth. Growth curves highlighted in the red box were modeled using slightly lower growth rates but assert growth.



Supplemental Figure 4. Growth levels. Growth level values for each condition. Blue horizontal line represents the threshold used for defining growth and no growth.

References

1. Aziz RK, Bartels D, Best AA, *et al.*: **The RAST Server: rapid annotations using subsystems technology.** *BMC Genomics.* 2008; **9**: 75.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Overbeek R, Begley T, Butler RM, *et al.*: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.** *Nucleic Acids Res.* 2005; **33**(17): 5691–5702.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Oberhardt MA, Puchalka J, Fryer KE, *et al.*: **Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1.** *J Bacteriol.* 2008; **190**(8): 2790–2803.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Orth JD, Thiele I, Palsson B: **What is flux balance analysis?** *Nat Biotechnol.* 2010; **28**(3): 245–248.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Schilling CH, Letscher D, Palsson BO: **Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective.** *J Theor Biol.* 2000; **203**(3): 229–248.
PubMed Abstract | Publisher Full Text
6. Raman K, Chandra N: **Flux balance analysis of biological systems: applications and challenges.** *Brief Bioinform.* 2009; **10**(4): 435–49.
PubMed Abstract | Publisher Full Text
7. Keseler IM, Bonavides-Martinez C, Collado-Vides J, *et al.*: **EcoCyc: a comprehensive view of Escherichia coli biology.** *Nucleic Acids Res.* 2009; **37**(Database issue): D464–D470.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Bochner BR: **New technologies to assess genotype-phenotype relationships.** *Nat Rev Genet.* 2003; **4**(4): 309–14.
PubMed Abstract | Publisher Full Text
9. Johnson DA, Tetu SG, Phillippy K, *et al.*: **High-throughput phenotypic characterization of Pseudomonas aeruginosa membrane transport genes.** *PLoS Genet.* 2008; **4**(10): e1000211.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Perkins AE, Nicholson WL: **Uncovering new metabolic capabilities of Bacillus subtilis using phenotype profiling of rifampin-resistant rpoB mutants.** *J Bacteriol.* 2008; **190**(3): 807–814.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Prüss BM, Campbell JW, Van Dyk TK, *et al.*: **FlhD/FlhC Is a regulator of anaerobic respiration and the Entner-Doudoroff pathway through Induction of the methyl-accepting chemotaxis protein Aer.** *J Bacteriol.* 2003; **185**(2): 534–543.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Viti C, Decorosi F, Mini A, *et al.*: **Involvement of the oscA gene in the sulphur starvation response and in Cr(VI) resistance in Pseudomonas corrugata 28.** *Microbiology.* 2009; **155**(Pt 1): 95–105.
PubMed Abstract | Publisher Full Text
13. Zhou L, Lei XH, Bochner BR, *et al.*: **Phenotype microarray analysis of *Escherichia coli* K-12 mutants with deletions of all two-component systems.** *J Bacteriol.* 2003; **185**(16): 4956–4972.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Mols M, de Been M, Zwietering MH, *et al.*: **Metabolic capacity of Bacillus cereus strains ATCC 14579 and ATCC 10987 interlinked with comparative genomics.** *Environ Microbiol.* 2007; **9**(12): 2933–2944.
PubMed Abstract | Publisher Full Text
15. Borglin S, Joyner D, DeAngelis KM, *et al.*: **Application of phenotypic microarrays to environmental microbiology.** *Curr Opin Biotechnol.* 2012; **23**(1): 41–48.
PubMed Abstract | Publisher Full Text
16. Covert MW, Knight EM, Reed JL, *et al.*: **Integrating high-throughput and computational data elucidates bacterial networks.** *Nature.* 2004; **429**(6987): 92–96.
PubMed Abstract | Publisher Full Text
17. Feist AM, Henry CS, Reed JL, *et al.*: **A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.** *Mol Syst Biol.* 2007; **3**: 121.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Oh YK, Palsson BO, Park SM, *et al.*: **Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data.** *J Biol Chem.* 2007; **282**(39): 28791–28799.
PubMed Abstract | Publisher Full Text
19. Bochner B, Gomez V, Ziman M, *et al.*: **Phenotype microArray profiling of Zymomonas mobilis ZM4.** *Appl Biochem Biotechnol.* 2010; **161**(1–8): 116–123.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Line JE, Hiett KL, Guard-Bouldin J, *et al.*: **Differential carbon source utilization by Campylobacter jejuni 11168 in response to growth temperature variation.** *J Microbiol Methods.* 2010; **80**(2): 198–202.
PubMed Abstract | Publisher Full Text
21. Neidhardt FC, Bloch PL, Smith DF: **Culture medium for enterobacteria.** *J Bacteriol.* 1974; **119**(3): 736–747.
PubMed Abstract | Free Full Text
22. Monod J: **The Growth of Bacterial Cultures.** *Annu Rev Microbiol.* 1949; **3**: 371–394.
Publisher Full Text
23. Zwietering MH, Jongenburger I, Rombouts FM, *et al.*: **Modeling of the bacterial growth curve.** *Appl Environ Microbiol.* 1990; **56**(6): 1875–1881.
PubMed Abstract | Free Full Text
24. Jones E, Oliphant T, Peterson P: **SciPy: Open source scientific tools for Python.** 2001.
Reference Source
25. **Conjugate Gradient Methods.** *Numerical Optimization. Springer Series in Operations Research and Financial Engineering.* Springer New York. 2006; 101–134.
Publisher Full Text
26. Edwards RA, Haggerty JM, Cassman N, *et al.*: **Microbes, metagenomes and marine mammals: enabling the next generation of scientist to enter the genomic era.** *BMC Genomics.* 2013; **14**: 600.
PubMed Abstract | Publisher Full Text | Free Full Text
27. Altschul SF, Madden TL, Schäffer AA, *et al.*: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res.* 1997; **25**(17): 3389–3402.
PubMed Abstract | Publisher Full Text | Free Full Text
28. Mitchell NB, Levine M: **Nitrogen Availability as an Aid in the Differentiation of Bacteria in the Coli-Aerogenes Group.** *J Bacteriol.* 1938; **36**(6): 587–598.
PubMed Abstract | Free Full Text
29. Kim J, Reed JL: **Refining metabolic models and accounting for regulatory effects.** *Curr Opin Biotechnol.* 2014; **29**: 34–38.
PubMed Abstract | Publisher Full Text
30. Henry CS, DeJongh M, Best AA, *et al.*: **High-throughput generation, optimization and analysis of genome-scale metabolic models.** *Nat Biotechnol.* 2010; **28**(9): 977–982.
PubMed Abstract | Publisher Full Text
31. Orth JD, Conrad TM, Na J, *et al.*: **A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism--2011.** *Mol Syst Biol.* 2011; **7**: 535.
PubMed Abstract | Publisher Full Text | Free Full Text
32. Cuevas DA, Garza D, Sanchez SE, *et al.*: **Phenotypic profiling data for elucidating genomic gaps.** *Figshare.* 2016.
Data Source
33. Cuevas DA, Garza D, Sanchez SE, *et al.*: **PMAnalyzer.** *Zenodo.* 2014.
Data Source
34. O'Brien EJ, Lerman JA, Chang RL, *et al.*: **Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction.** *Mol Syst Biol.* 2013; **9**(1): 693.
PubMed Abstract | Publisher Full Text | Free Full Text
35. O'Brien EJ, Palsson BO: **Computing the functional proteome: recent progress and future prospects for genome-scale models.** *Curr Opin Biotechnol.* 2015; **34**: 125–34.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ? ✓

Version 2

Reviewer Report 09 December 2016

https://doi.org/10.5256/f1000research.10053.r17041

✓ **Aaron Best**
Department of Biology, Hope College, Holland, MI, USA

The authors have adequately addressed concerns raised in the initial review.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 27 March 2015

https://doi.org/10.5256/f1000research.5480.r7178

? **Aaron Best**
Department of Biology, Hope College, Holland, MI, USA

The article by Cuevas and colleagues represents an important attempt to streamline the assessment of phenotypic growth data in the context of genome scale metabolic models. The approach incorporates a non-proprietary phenotyping assay that should be accessible to a wide variety of research groups. The data are then coupled to a developing genome analysis and modeling environment, KBase and associated systems (the SEED and RAST) to enable evaluation and iterative refinement of metabolic models. The primary contribution of the work is to provide a

methodological path for acquisition and reasonably quick integration of these data with genomic information.

Below, I provide a series of questions, concerns and suggestions that I have for the each section of the manuscript as written.

Introduction

This section is generally well written, though I think the context for this type of analysis could be better described... that is, beyond mentioning simply KBase and RAST. Perhaps providing rationale for potential advantages of using this combination of systems rather than other options would be useful for the reader.

In the final paragraph, the term "ground truth" is used to describe the phenotype data with respect to the metabolic modeling that will be performed. Foreshadowing some comments later, what is the evidence that the phenotype data are actually correct for all conditions tested?

Methods

First, thank you for the detailed description of the methods. These are clearly written overall. Some detailed comments/suggestions.

Describing growth of the initial cells and cultures from glycerol stocks, please define the shaking parameter (rpm) and define "agitation". These parameters can be critical factors for reproduction of experiments and are often organism dependent.

Second full paragraph, you have defined the contents of 1X MOPS, but you also define it two paragraphs later in the context of the recipe for the basal media. This first instance could be removed and simply refer to the recipe later described.

Question: In the plate format for growing the cells, you indicate that plates are sealed with a PCR grade plate film. What does this do to the aerobic/anaerobic state of each well? Is there any opportunity for gas exchange during incubation? Also, is there any shaking going on during incubation on the plate reader? It might be worth mentioning about caveats of usage of carbon/nitrogen sources being limited to these conditions, which aren't exactly known. **Also, the modeling could be impacted by the aerobic/anaerobic status of the environment. Was modeling performed under both conditions? Would this impact the accuracy of the modeling results?**

I note the storage condition was at room temperature. Are all of the substrates stable at room temperature? How long would each stock be stored prior to use for replicates?

In the "Sequencing and metabolic reconstruction..." section:

1. The sentence that starts with "FBA was used to determine...". This is a confusing sentence. What is meant by this?

2. The KBase workspace, Citrobacter_sedlakii_119, does not appear to exist in the current public release of KBase (as of March 25, 2015). I also am unable to find any FBA model

objects searching for various forms of *Citrobacter* and *sedlakii*. There are not an public narratives that would match the series of commands that you describe as being run and as freely accessible. **This needs to be corrected, likely by building a public narrative in the current system.**

3. Are the named commands for KBase still valid in the current production version of the system? It would be useful to include what apps and methods correspond to these commands. **A public narrative in the current version of KBase would make this study replicable and easily transferred to other model systems.**

The github page for the PMAnalyzer software is good... to the point, clear.

The explanation of the logistic model and absorbance data is also clear. In the description of the *growth* value, you end by stating that this is boiled down to a boolean growth/no growth status for each condition. I understand why this is done, given that the model reconciliation with growth phenotypes is occurring on a boolean level, but how much information is being lost by making this experimental design decision? The nature of the growth can be very important for understanding how the organism is behaving in an environment. The more immediate consequence of this decision is in the interpretation of False Negatives by the model (where the phenotype assay says "growth" and the model says "no growth"). **How many of the false negatives had *growth* values near the 0.5 cutoff? The allantoin example could be a case like this (*growth* = 0.529, from curve_logistic_parameters.csv). The growth curve asymptote appears to be near 0.25 (Fig. 2a). This is very similar to values that are considered "no growth" phenotypes. Does it make sense to have the model gap fill three reactions in this case? Table 1 might be made more complete by adding a column for the *growth* value for each condition rather than that sitting in the supplemental data files (alternatively, highlight in the text that these values are given in that file). Related to this, in Supplemental Figure 3, you could highlight the point that represents allantoin. It would also be useful to highlight the water, negative control in the Supp. Fig. 3.**

This brings me to questions about how confidence in the *growth* value is determined (if at all). I see that the median value of the biological replicates is used to determine the y logistic and thus the *growth* value. I also see that standard error is indicated in Fig. 2a graphs. However, this does not allow for statistical evaluation of the *growth*. **Would it be better to calculate *growth* for each replicate independently and then determine an average *growth* value with error around these? Perhaps there is a better statistical approach. In any case, this comes back to being able to state some confidence in these values to aid interpretation of potentially borderline cases.** Please define "sse" in the curve_logistic_parameters.csv file.

In the "RAST annotations" section, last paragraph. How does this fit in with the gap filling process for the model? Is the context information in close genomes actually used in the gap fill process, or is it a post hoc attribution of higher confidence to the gap fills that are included in the model?

Results

The statistics on the genome assembly are worse than I would expect to see. In particular, is the coverage based on alignment by blastn to *C. koserii* a reasonable number? I can't quickly evaluate if this is typical of different *Citrobacter* genomes. How does a low coverage (~70%) affect the outcome of presence/absence of genes in the annotation and subsequent modeling process. In

reading the results, it appears that the majority of reactions in the network are identified, but it may be worth addressing this explicitly.

I note that you used manual inspection of growth curves just under the 0.5 cutoff... this is another area in which a statistical confidence in that value might help. If this is to become truly high throughput, manual inspection becomes untenable except in a few cases.

In the description of gap filling reactions for complex media, you mention that EC 2.2.1.7 was not found, but is likely due to a frameshift error. Was any follow up sequencing or PCR performed to confirm the error or presence/absence of the gene in *C. sedlakii*? Or even just a blastx analysis of the region? **In general, when you are discussing the evaluation of annotations in RAST, figures (supplemental) of key genomic regions would help the reader to evaluate the statements being made.**

In the paragraph beginning with, "Using the base model, the 90 well simulation resulted in...", you have a sentence that starts with, "Note:". This is a confusing sentence and structure. What are you trying to point out here? What reactions are being referred to?

What percentage of gap filled reactions are transport reactions? Stating this clearly would improve clarity.

The last statement in the Results section focuses on false positive conditions. **Do you have any thoughts as to why these are coming up as FP? There is no follow up in the discussion about this. Are they central in a network of reactions, are they dual use reactions, etc.?**

Discussion

RAST annotations and gap-filled reactions section:

This section would also benefit from a supplemental figure that serves as an example of what is being discussed (also mentioned above).

What is the connection between KBase, RAST and SEED? How does updating in one affect the others? This question gets at an assumption in the text that the relationships among systems are known to the reader. The text could be clarified, or key references added.

FBA false positives section:

Please expand to include more specific discussion of the 6 FP reactions identified at the end of the Results section. What are the *growth* values for these? Are any of them borderline?

Last paragraph:

It would be good to quantify what “several” means with respect to the number of metabolic pathways being targeted.

“...available in a day of using RAST and KBase.” This sentence implies that sequencing, annotation, and model reconstruction can happen in a single day. This should refer only to the use of

sequence data. Also, there in no mention of the phenotype data here in this context. I think it would be better to highlight that the system allows the user to produce a reasonably robust metabolic model quickly, giving more opportunity for in depth analysis of discrepancies and manual curation of the model given the phenotypic data.

What is the link to the web service for the PMAnalyzer?

Points to Address

I've bolded several items in the above format for this review that I would consider to be major points to address and would make the manuscript stronger. Given that this is a methods paper, it is imperative that others can reproduce the work and/or employ the approach in other organism systems. Please update methods as requested above, paying particular attention to the KBase functionality and workflow.

Competing Interests: Non-financial competing interests include: I have worked as a collaborator with several of the authors on projects related to the SEED, RAST and KBase. I have not worked directly with the lead author. Financial competing interests: None to declare.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 16 Sep 2016

Daniel Cuevas, San Diego State University, San Diego, USA

In the final paragraph, the term "ground truth" is used to describe the phenotype data with respect to the metabolic modeling that will be performed. Foreshadowing some comments later, what is the evidence that the phenotype data are actually correct for all conditions tested?

Response - The 96-well microtiter plate technology has been established and used in many research studies to date, some of which have been cited in this manuscript; thus, the efficacy of the data has been verified. The phenotype response taken from these growth curves is identifying growth or no growth of the bacteria in specific minimal media conditions. Classifying growth for a given sample can be difficult because there is not an established rule that translates quantitative optical density measurements into a qualitative growth/ no growth response. When a growth curve displays an ambiguous phenotype, such as described in the manuscript with the allantoin-based condition, other laboratory techniques can be used to offer a more precise answer.

- **Methods**

Describing growth of the initial cells and cultures from glycerol stocks, please define the shaking parameter (rpm) and define "agitation". These parameters can be critical factors for reproduction of experiments and are often organism dependent.

Response - The 250 rpm shaking speed has been added to the manuscript.

In the plate format for growing the cells, you indicate that plates are sealed with a PCR grade plate film. What does this do to the aerobic/anaerobic state of each well? Is there any opportunity for gas exchange during incubation? Also, is there any shaking going on during incubation on the plate reader? It might be worth mentioning about caveats of usage of carbon/nitrogen sources being limited to these conditions, which aren't exactly known.

Response - The PCR grade plate film still allows gas exchange to occur, and shaking does occur on the plate reader. This has been clarified in the methods section of the manuscript.

Also, the modeling could be impacted by the aerobic/anaerobic status of the environment. Was modeling performed under both conditions? Would this impact the accuracy of the modeling results?

Response - Flux-balance analysis was performed with oxygen exchange occurring in the metabolic model.

I note the storage condition was at room temperature. Are all of the substrates stable at room temperature? How long would each stock be stored prior to use for replicates?

Response - Yes, all substrates are stable at room temperature. Substrate stocks were prepared on a weekly basis.

The sentence that starts with "FBA was used to determine...". This is a confusing sentence. What is meant by this?

Response - Flux-balance analysis answers the question: with the given genome-scale metabolic model and the nutrients present in the environment, does the genome-scale metabolic model contain biochemical reactions that will intake the nutrients and create the necessary biomass components for cellular growth? Thus, FBA was used here to determine if the model is capable of growth in the same conditions as the MAPs.

The KBase workspace, Citrobacter_sedlakii_119, does not appear to exist in the current public release of KBase (as of March 25, 2015). I also am unable to find any FBA model objects searching for various forms of Citrobacter and sedlakii. There are not any public narratives that would match the series of commands that you describe as being run and as freely accessible. This needs to be corrected, likely by building a public narrative in the current system.

Response - The SBML files for the draft model and gap-filled models have been provided as supplementary material.

Are the named commands for KBase still valid in the current production version of the system? It would be useful to include what apps and methods correspond to these commands. A public narrative in the current version of KBase would make this study replicable and easily transferred to other model systems.

Response - At this time KBase does not use the IRIS system to perform genome-scale metabolic modelling. KBase now uses the Narrative graphical workflow to perform the same functions using the same data types. KBase has released publicly available narratives that describe these workflows (e.g., https://narrative.kbase.us/#appcatalog/app/fba_tools/build_metabolic_model/release).

In the description of the growth value, you end by stating that this is boiled down to a boolean growth/no growth status for each condition. I understand why this is done, given that the model reconciliation with growth phenotypes is occurring on a boolean level, but how much information is being lost by making this experimental design decision? The nature of the growth can be very important for understanding how the organism is behaving in an environment. The more immediate consequence of this decision is in the interpretation of False Negatives by the model (where the phenotype assay says "growth" and the model says "no growth"). How many of the false negatives had growth values near the 0.5 cutoff? The allantoin example could be a case like this (growth = 0.529, from curve_logistic_parameters.csv). The growth curve asymptote appears to be near 0.25 (Fig. 2a). This is very similar to values that are considered "no growth" phenotypes. Does it make sense to have the model gap fill three reactions in this case? Table 1 might be made more complete by adding a column for the growth value for each condition rather than that sitting in the supplemental data files (alternatively, highlight in the text that these values are given in that file). Related to this, in Supplemental Figure 3, you could highlight the point that represents allantoin. It would also be useful to highlight the water, negative control in the Supp. Fig. 3.

Response - Supplementary Figure 4 has been generated to show the different growth levels in terms of the 0.5 growth level cutoff.

Would it be better to calculate growth for each replicate independently and then determine an average growth value with error around these? Perhaps there is a better statistical approach. In any case, this comes back to being able to state some confidence in these values to aid interpretation of potentially borderline cases.

Response - Yes, this does indeed provide some statistical evidence of the growth level and each of the other growth parameters. Although the updated PMAnalyzer pipeline now does this order of analysis, it did not affect the results to this experiment; thus, the results were not altered in terms of identifying growth and no growth conditions.

Please define "sse" in the curve_logistic_parameters.csv file.

Response - “SSE” refers to the sum-squared error calculated between the logistic fitted

growth model and the OD measurements.

In the "RAST annotations" section, last paragraph. How does this fit in with the gap filling process for the model? Is the context information in close genomes actually used in the gap fill process, or is it a post hoc attribution of higher confidence to the gap fills that are included in the model?

Response - This refers to post hoc, manual efforts made after the gap-filling process.

- **Results**

The statistics on the genome assembly are worse than I would expect to see. In particular, is the coverage based on alignment by blastn to C. koserii a reasonable number? I can't quickly evaluate if this is typical of different Citrobacter genomes. How does a low coverage (~70%) affect the outcome of presence/absence of genes in the annotation and subsequent modeling process. In reading the results, it appears that the majority of reactions in the network are identified, but it may be worth addressing this explicitly.

Response - The genome alignment to *C. koseri* were meant to paint a picture of the similarity of its DNA sequence to the *C. sedlakii*, which might lend insight into why some of the genes were not identified with a functional role. This affects the presence of functional roles as many of those putative genes are not assigned any role, thus leaving gaps in our metabolic model.

In the paragraph beginning with, "Using the base model, the 90 well simulation resulted in...", you have a sentence that starts with, "Note:". This is a confusing sentence and structure. What are you trying to point out here? What reactions are being referred to?

Response - Here I am pointing out that through gap-filling for only the 13 false negative conditions listed in Table 2, the other 35 false negative conditions were corrected, i.e., all 48 conditions where FBA asserted false negative results now assert true positive results. This has been clarified in the updated manuscript.

What percentage of gap filled reactions are transport reactions? Stating this clearly would improve clarity.

Response - This 46% has been added to the manuscript.

- **Discussion**

What is the connection between KBase, RAST and SEED? How does updating in one affect the others? This question gets at an assumption in the text that the relationships among systems are known to the reader. The text could be clarified, or key references added.

Response - The references for RAST and the SEED database explain their relationships. RAST uses the SEED subsystems information to annotate genomic sequences.

Please expand to include more specific discussion of the 6 FP reactions identified at the end of the Results section. What are the growth values for these? Are any of them borderline?

Response - The issue has been addressed in the recent changes. Clarifications and explanations have been included in the Results and Discussion sections.

"...available in a day of using RAST and KBase." This sentence implies that sequencing, annotation, and model reconstruction can happen in a single day. This should refer only to the use of sequence data. Also, there in no mention of the phenotype data here in this context. I think it would be better to highlight that the system allows the user to produce a reasonably robust metabolic model quickly, giving more opportunity for in depth analysis of discrepancies and manual curation of the model given the phenotypic data.

Response - This clarification has been made in the manuscript.

What is the link to the web service for the PMAnalyzer?

Response- The link (https://vdm.sdsu.edu/pmanalyzer) has been added to the manuscript.

Competing Interests: No competing interests were disclosed.

Reviewer Report 30 September 2014

https://doi.org/10.5256/f1000research.5480.r6026

?

Matthew A. Oberhardt
School of Computer Sciences & Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel

Genome-scale metabolic modeling (GSM) has gained interest over the last 2 decades as an increasingly powerful methodology for interrogating cell physiology and function. One of the holy grails in this field is the development of high quality models fully automatically, which would enable analysis of organisms from sequencing data alone, and, for example, integration of modeling with the interpretation of metagenomic data. A large contributor towards this end is the group at Argonne National Labs, which has produced RAST (a genome sequencing and annotation database), SEED (an automated platform to build genome-scale metabolic models), and now Kbase (an integrated platform that allows users to utilize the power of RAST, SEED, and many other functionalities).

This article describes a new, automated pipeline for integrating phenotyping data (i.e., growth of an organism on many individual substrates in a microwell plate, such as in Biolog analysis) with genome-scale metabolic modeling in order to improve an automatically built genome-scale metabolic reconstruction produced through the automated RAST/Kbase system. This is an important problem and a technical issue that can greatly improve model reconstruction, automated and manual (the early stages of manual reconstruction are typically automated nowadays as well). The approach seems to provide satisfactory results. However, I have a few concerns about the approach, relating to its originality and some of its features. In general, it is not totally clear from the paper what its main contribution is, and how it is differentiated from previous work. This should be explicitly explained in the introduction, etc.

My specific criticisms are as follows:

Major criticisms:

1. As far as I can tell, the problem that PMAnalyzer solves has already been solved previously. For example, in the original paper describing SEED (Henry *et al.*, 2010), 22 models are automatically optimized against existing Biolog data. The paper reads: "A modified version of the Growmatch algorithm was included in the Model SEED pipeline to identify and correct the possible errors in the models that cause the incorrect predictions [in Biolog and essentiality data]..." In order to claim that PMAnalyzer is novel, it must be compared to such previous methods and shown to be superior or different in some way (I might simply not understand the difference; this should be explained clearly in the text, or some comparison shown).

2. In general (and as stated before), please compare this work to previous works and explain what is the main scientific novelty of the paper (i.e. how does PMAnalyzer differ from previous works? Also if the sequencing / analysis of this organism is novel, that should also be explicitly stated as well).

3. It is not clear from the text what part of the model building pipeline PMAnalyzer actually does (i.e., does it only analyze the growth curves? Does it do that and also run the gap filling? Etc.). Please explain this explicitly in the paper. I suggest also providing a schematic similar to Figure 1, but that specifically shows what the inputs and outputs to PMAnalyzer are (and optionally shows some of the internal mechanics of PMAnalyzer).

4. Please remark on and justify whether there is an optimization step to reduce false positives (which is mentioned in the discussion, but not the results). The final paragraph of the results lists false positives, but doesn't go into detail or attempt to explain why these occurred; I suggest that the authors give some explanations here if possible on why these are tricky biological cases.

5. Figure 3 would be greatly improved if the authors listed on each panel whether it was called 'growth' or 'non-growth' by PMAnalyzer. This is, after all, a way for the reader to visually validate the method.

Minor criticisms:

1. Please explain the genesis of equation 4, as it is unclear how/why it was formulated this way and it forms the critical cut-off criterion for the calls made in PMAnalyzer.

2. In the section 'RAST annotations', the authors state that ‘following gap-filling, all missing reactions … were cross-checked with the SEED to find similarly named reactions’ -- similar to what? Is this a comparison between databases held in SEED vs. in RAST? Please clarify this.

3. It would be interesting/informative to see Figure 2 in context of other models, especially those of close neighbors (e.g., *e. coli* and *C. koseri*). I suggest that the authors provide histograms of subsystems in one/both of those organism as well for comparison.

4. The authors mention in the Discussion that there is an automated web server for executing PMAnalyzer. However, I could not find the link. Can they please link to this or remove this sentence?

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 16 Sep 2016

Daniel Cuevas, San Diego State University, San Diego, USA

Major criticisms:

As far as I can tell, the problem that PMAnalyzer solves has already been solved previously. For example, in the original paper describing SEED (Henry et al., 2010), 22 models are automatically optimized against existing Biolog data. The paper reads: “A modified version of the Growmatch algorithm was included in the Model SEED pipeline to identify and correct the possible errors in the models that cause the incorrect predictions [in Biolog and essentiality data]…” In order to claim that PMAnalyzer is novel, it must be compared to such previous methods and shown to be superior or different in some way (I might simply not understand the difference; this should be explained clearly in the text, or some comparison shown).

Response - Biolog data has been more commonly used by others to describe phenotypic response in various media sources. These data can help enhance the accuracy of genome-scale metabolic reconstructions, as described by Henry *et al*, 2010. It is important to note that the Multi-phenotype Assay Plates (MAPs) described here are different from Biolog plates: MAPs are non-proprietary assays used to measure biomass accumulation whereas Biolog technology measure substrate utilization. However, the MAPs are serving a similar purpose for model reconciliation as Biolog plates do as described by the GrowMatch paper. The novelty claimed here is the method in combining the high-throughput MAPs technology, the PMAnalyzer pipeline, and the KBase modeling pipeline. PMAnalyzer quickly and automatically calculates growth profiles for a wide spectrum of sugars, amino acids, and other compounds which the KBase modeling environment can also simulate with FBA. In addition to the FBA, KBase is performing the gap-fill and gap-gen algorithms to reconcile the model. This process uses the modified version of the GrowMatch algorithm to identify those changes (e.g., adding reactions, making reactions reversible).

In general (and as stated before), please compare this work to previous works and explain what is the main scientific novelty of the paper (i.e. how does PMAnalyzer differ from previous works? Also if the sequencing / analysis of this organism is novel, that should also be explicitly stated as well).

Response - This issue has been addressed in the recent changes to the manuscript where we clarify the difference between the Multi-phenotype Assay Plates (MAPs) technology and the Biolog Phenotype MicroArray system, and have also been explained in the response to the previous comment.

It is not clear from the text what part of the model building pipeline PMAnalyzer actually does (i.e., does it only analyze the growth curves? Does it do that and also run the gap filling? Etc.). Please explain this explicitly in the paper. I suggest also providing a schematic similar to Figure 1, but that specifically shows what the inputs and outputs to PMAnalyzer are (and optionally shows some of the internal mechanics of PMAnalyzer).

Response - This issue has been addressed in the recent changes. An additional flowchart has been added to Figure 1 describing in further detail the workflow of PMAnalyzer.

Please remark on and justify whether there is an optimization step to reduce false positives (which is mentioned in the discussion, but not the results). The final paragraph of the results lists false positives, but doesn't go into detail or attempt to explain why these occurred; I suggest that the authors give some explanations here if possible on why these are tricky biological cases.

Response - The issue has been addressed in the recent changes. Clarifications and explanations have been included in the Results and Discussion sections.

Figure 3 would be greatly improved if the authors listed on each panel whether it was called 'growth' or 'non-growth' by PMAnalyzer. This is, after all, a way for the reader to visually validate the method.

Response - Figure 3 has been updated to show Growth and No Growth curves.

Minor criticisms:

Please explain the genesis of equation 4, as it is unclear how/why it was formulated this way and it forms the critical cut-off criterion for the calls made in PMAnalyzer.

Response - Equation 4 is a type of arithmetic mean that is least prone to noise in the data. Originally, the data input into this equation was the raw OD 600nm measurements; however, after fitting the logistic model, empirical data showed that using the fitted values

and introducing the asymptotic value into the equation was able to further separate those growth curves displaying growth.

In the section 'RAST annotations', the authors state that 'following gap-filling, all missing reactions ... were cross-checked with the SEED to find similarly named reactions' -- similar to what? Is this a comparison between databases held in SEED vs. in RAST? Please clarify this.

Response - This issue has been addressed in the recent changes with clarifications and explanations.

It would be interesting/informative to see Figure 2 in context of other models, especially those of close neighbors (e.g., e. coli and C. koseri). I suggest that the authors provide histograms of subsystems in one/both of those organism as well for comparison.

Response - Figure 2 has been updated to provide this information.

The authors mention in the Discussion that there is an automated web server for executing PMAnalyzer. However, I could not find the link. Can they please link to this or remove this sentence?

Response - The link (https://vdm.sdsu.edu/pmanalyzer) has been added to the manuscript.

Competing Interests: No competing interests were disclosed.