Title

Model-driven discovery of underground metabolic functions in Escherichia coli.

Journal

Proceedings of the National Academy of Sciences of USA, 112(3)

Authors

Guzmán, Gabriela
Utrilla, José
Nurk, Sergey
et al.

Publication Date

2015-01-20

DOI

10.1073/pnas.1414218112

Peer reviewed

Model-driven discovery of underground metabolic functions in *Escherichia coli*

Gabriela I. Guzmán[a], José Utrilla[a], Sergey Nurk[b], Elizabeth Brunk[c], Jonathan M. Monk[d], Ali Ebrahim[a], Bernhard O. Palsson[a,e,f], and Adam M. Feist[a,e,1]

Departments of [a]Bioengineering, [d]NanoEngineering, and [f]Pediatrics, University of California, San Diego, La Jolla, CA 92093; [b]Algorithmic Biology Laboratory, St. Petersburg Academic University, Russian Academy of Sciences, St. Petersburg, Russia; [c]Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Emeryville, CA 94608; and [e]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark

Edited by Marc W. Kirschner, Harvard Medical School, Boston, MA, and approved December 5, 2014 (received for review August 5, 2014)

Enzyme promiscuity toward substrates has been discussed in evolutionary terms as providing the flexibility to adapt to novel environments. In the present work, we describe an approach toward exploring such enzyme promiscuity in the space of a metabolic network. This approach leverages genome-scale models, which have been widely used for predicting growth phenotypes in various environments or following a genetic perturbation; however, these predictions occasionally fail. Failed predictions of gene essentiality offer an opportunity for targeting biological discovery, suggesting the presence of unknown underground pathways stemming from enzymatic cross-reactivity. We demonstrate a workflow that couples constraint-based modeling and bioinformatic tools with KO strain analysis and adaptive laboratory evolution for the purpose of predicting promiscuity at the genome scale. Three cases of genes that are incorrectly predicted as essential in *Escherichia coli*—*aspC*, *argD*, and *gltA*—are examined, and isozyme functions are uncovered for each to a different extent. Seven isozyme functions based on genetic and transcriptional evidence are suggested between the genes *aspC* and *tyrB*, *argD* and *astC*, *gabT* and *puuE*, and *gltA* and *prpC*. This study demonstrates how a targeted model-driven approach to discovery can systematically fill knowledge gaps, characterize underground metabolism, and elucidate regulatory mechanisms of adaptation in response to gene KO perturbations.

underground metabolism | substrate promiscuity | systems biology | isozyme discovery | genome-scale modeling

The notion that enzymes are highly specialized to carry out a single function is often untrue. It has been demonstrated that many enzymes exhibit flexibility, or promiscuity, in regard to what substrates their catalytic pockets recognize. This lack of substrate specificity can lead to accuracy-rate tradeoffs that may affect evolutionary trajectories (1). How has enzyme promiscuity shaped the evolution and divergence of organisms? The "patchwork" model theorizes that primitive enzymes possessed a high degree of substrate promiscuity because it conferred a greater degree of catalytic versatility when the pool of available enzymes was limited (2–5). The existence of promiscuous proteins further serves as a starting point for evolving new functions, allowing for novel adaptations. Thus, organisms may exhibit latent, underground metabolic pathways that form the basis of their capacity to adapt to changing environments (6–8). Substrate promiscuity, also referred to as "moonlighting activity" and "cross-reactivity," has thus been studied in terms of evolution, and ties have been made between enzymes and their superfamilies (9). How novel enzyme functions arise within superfamilies is thus examined, and provides a basis for predicting promiscuous behavior among these protein families. However, defining targets for studies of promiscuity outside of these families and on a larger scale can become quite challenging.

Enzyme promiscuity has become widely accepted and examined on the enzyme level from a biochemical standpoint (10). These detailed biochemical studies provide an in vitro view of enzyme promiscuity and may be extended to reflect the promiscuity of other proteins based on sequence homology or enzyme familial relationships. In the present study, this task is approached from a different perspective by taking advantage of in vivo experimental techniques to gain insight into activities that are more physiologically relevant. In this way, as has been demonstrated in other in vivo studies, many of the challenges associated with removing enzymes from their native environment are circumvented (8). Specifically, the present study focuses on examination of the regulatory and evolutionary capacity of a cell in vivo. Theories regarding genome duplications have suggested that an enzyme with a side activity that is selected for may be enhanced via gene duplication followed by mutation accumulation (11). Thus, laboratory evolutions may provide insight into these evolutionary mechanisms involving enzyme promiscuity. Furthermore, exploration of an underground metabolic network that takes advantage of enzyme cross-reactivity through native regulatory adaptations is best examined in the context of a whole cell (6, 8).

A top-down, model-driven approach coupled with in vivo experimentation to explore enzyme promiscuity could provide new insights into the physiological role of underground metabolism and complement the current approaches to enzyme research. Computational predictions of gene essentiality are a commonly used application of genome-scale models and constraint-based modeling (12, 13). When these models fail to predict gene essentiality, it signifies a missing link in our knowledge of metabolism and provides targets for further exploration (14). Various computational algorithms, including SMILEY and GrowMatch, have been published with the intent of reconciling such knowledge gaps (15, 16). The following is a proof-of-principle study that

Significance

Organisms have evolved to take advantage of their environment. Enzymes drive this adaptability by displaying flexibility in terms of substrate specificity and catalytic promiscuity. This enzyme promiscuity has been observed in a limited number of laboratory experiments; however, a larger underground network of reactions may occur within a cell below the level of detection. It is not until a cell's metabolic capabilities are probed that these novel functions come to light. In this study, a workflow is presented for probing promiscuous activity at the genome scale. This workflow combines genome-scale reconstructions of metabolic networks with gene KOs and adaptive laboratory evolution. Such tools become increasingly important when designing drugs targeting pathogenic bacteria or engineering enzymes and bacteria for biotechnology applications.

Author contributions: G.I.G., J.U., B.O.P., and A.M.F. designed research; G.I.G., J.U., S.N., E.B., and J.M.M. performed research; G.I.G., S.N., E.B., J.M.M., and A.E. contributed new reagents/analytic tools; G.I.G., J.U., S.N., E.B., and A.M.F. analyzed data; and G.I.G. and A.M.F. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: afeist@ucsd.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1414218112/-/DCSupplemental.

demonstrates the advantages of a workflow for examining promiscuity at the genome scale that also encompasses an adaptive laboratory evolution (ALE) framework. Three cases are explored to illustrate the capabilities of such a targeted, top-down approach to uncover the underground, latent activities of enzymes that reconcile gaps in our knowledge of metabolism.

Results and Discussion

Developing a Model-Driven Workflow for Isozyme Discovery. The results from this study demonstrated that a top-down systems approach could be used to drive the discovery of enzyme substrate promiscuity by using three genes, *aspC*, *argD*, and *gltA*, that were incorrectly identified to be essential as inputs. The isozyme discovery workflow presented in this study is a prime example of targeted analysis based on systems-level insights (Fig. 1).

The first step in the isozyme discovery workflow was to identify the targets for exploration. These targets come from performing flux balance analysis (FBA) gene essentiality simulations in *Escherichia coli* using the *i*JO1366 metabolic reconstruction (17, 18). When discussing computational gene essentiality predictions, the term "false-negative prediction" refers to a situation in which a gene is predicted to be essential but is experimentally observed to be nonessential. This type of prediction failure can stem from lack of knowledge of an alternate pathway or isozyme (14). All genes associated with false-negative predictions in *i*JO1366 were identified, and those genes with high-confidence candidate isozymes, based on sequence homology, were used as examples for this study. To identify potential isozymes based on sequence homology, the National Center for Biotechnology Information's BLASTp algorithm (19) was run for each protein sequence (results are summarized in Table S1). An expect (E) value of <1E-40 and high-sequence identity percentage were used as cutoffs for candidates.

Following identification of false-negative targets and potential isozymes, experiments were conducted to determine which isozyme candidate might explain the modeling failure. The KO strain corresponding to the false-negative target was examined. Growth on glucose minimal medium of the KO strain was confirmed. It was then hypothesized that an isozyme was compensating for the lost function of the primary gene that was knocked out. This hypothesis was tested by exploring expression of the putative isozymes in the primary KO strain. RT-quantitative PCR (qPCR) analysis was performed, and if an isozyme candidate was up-regulated, the next step in the workflow was followed (Fig. 1).

Following confirmation of up-regulation of the candidate isozyme, a double-KO (DKO) strain was constructed. It was thus hypothesized that removal of the up-regulated isozyme candidate would lead to a synthetic lethal interaction if there remained no other isozymes. The next step in the workflow was to test the growth of the DKO strain and confirm a synthetic lethal interaction. If a synthetic lethal interaction was verified following at least 1 wk of incubation, then the isozyme was deemed to be correctly identified based on genetic and transcriptional evidence. A possible deviation from the above steps is also taken into consideration in this study. For example, if a DKO strain was not lethal, the possibility of an alternate isozyme was explored. The following sections describe the specific workflows followed for the three false-negative cases examined in this study (Fig. S1).

Implementation of this workflow resulted in three main findings: (*i*) discovery of seven isozyme associations and adaptive regulatory mechanisms for partially characterized enzymes in *E. coli*; (*ii*) development of an enzyme substrate promiscuity discovery tool that can easily be extended to fill other knowledge gaps in *E. coli*, as well as in other organisms; and (*iii*) the establishment of a more rigorous assessment of lethality with longer growth incubations to prevent false statements of lethality, particularly for high-throughput screens.

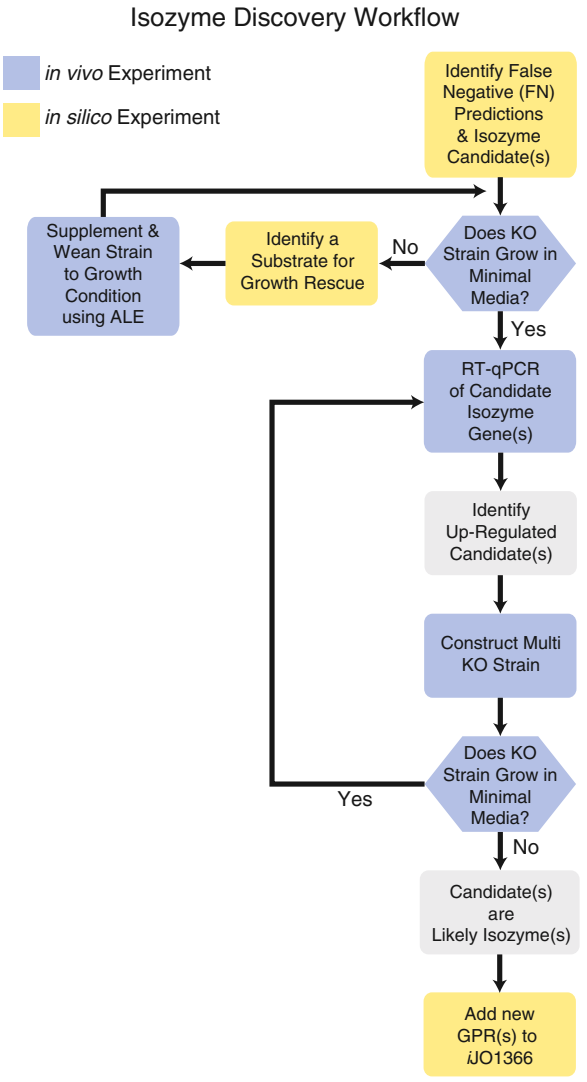

Fig. 1. Schematic of the general workflow used for isozyme discovery involving both in vivo and in silico experiments. Starting from the topmost box, false-negative model predictions and isozyme candidates were identified using FBA and BLASTp. The workflow was then followed vertically downward, examining KO strain growth, expression levels of candidate isozyme genes, and multi-KO strain phenotypes. Deviations from the schematic occurred when growth discrepancies were encountered. The workflow was terminated once a synthetic lethal interaction of a false-negative gene and isozyme candidate(s) was identified. The output was enzymatic activities characterized and added to the current genome-scale model reconstruction of *E. coli*.

Case 1: *aspC*–Asp Aminotransferase. The Asp aminotransferase in *E. coli*, encoded by the gene *aspC*, has been characterized as a broad-substrate, multifunctional enzyme that catalyzes the formation of Asp, Phe, and Tyr (20). Early studies have drawn links to the overlapping functions of the aminotransferases encoded by *aspC*, *tyrB*, and *ilvE* (21). The aromatic aminotransferase encoded by *tyrB* has shown activity in the synthesis of Phe, Tyr, and Leu (22), whereas the branched-chain aminotransferase, encoded by *ilvE*, has been associated with the synthesis of Ile, Leu, and Val (23). Previous studies reported that *aspC* KO strains are viable in both rich media and glucose minimal medium (24); however, *i*JO1366 model simulations predicted no growth on minimal medium. Given this false-negative prediction, BLASTp was then used and results pointed to *tyrB* as an isozyme candidate (Table S1).

Initial growth tests were performed to verify reports of nonessentiality. The growth data for the Δ*aspC* strain are illustrated in Fig. 2*B*. Growth of the Δ*tyrB* strain was also validated in this study. Following completion of initial growth characterizations, RT-qPCR analysis of the *tyrB* isozyme target was performed in the Δ*aspC* and WT strains. qPCR analysis showed up-regulation of *tyrB* in the Δ*aspC* strain, with a fold change of 4.7 compared with the WT strain (Fig. 2*A*). Therefore, the next step in the workflow was followed, and construction of the Δ*aspC*Δ*tyrB* DKO strain was performed. Growth of the Δ*aspC*Δ*tyrB* strain was monitored (Fig. 2*B*) and the Δ*aspC*Δ*tyrB* KO pair was deemed synthetically lethal based on this genetic evidence. Therefore, successful execution of the workflow identified the isozyme link between *aspC* and *tyrB*.

A secondary result of executing this method was the discovery of an association between *ilvE* and L-Tyr biosynthesis. Efforts were placed on finding amino acid supplements that would enable growth of the Δ*aspC*Δ*tyrB* strain to validate the functions of this interrelated trio of genes further. Growth characterizations were performed using various combinations of amino acid supplementation, including L-Asp, L-Tyr, L-Phe, and L-Leu. Gene KO simulations of growth on glucose plus supplementation of all combinations of the *aspC*- and *tyrB*-associated amino acids resulted in an expected requirement of Asp and Tyr for growth rescue of the Δ*aspC*Δ*tyrB* strain (Table S2). Experimental observations, however, showed that only Asp was required for growth rescue. It was therefore speculated that the aminotransferase encoded by *ilvE* was fulfilling the role of Tyr synthesis. The enzyme encoded by *ilvE* has shown some, although minimal, specific activity with Phe and Tyr in an in vitro assay (23). Thus, the overlapping in vivo functionality of these aminotransferases, *aspC*, *tyrB*, and *ilvE*, appeared to be greater than previously expected.

Finally, to examine the possibility of acquired mutations in the strains constructed in this study, genome resequencing was performed for the Δ*aspC* and Δ*tyrB* strains (25). A summary of the output for these strains is shown in Table S3. The Δ*aspC* and Δ*tyrB* strains showed no apparent mutations in the coding regions of the related isozymes examined.

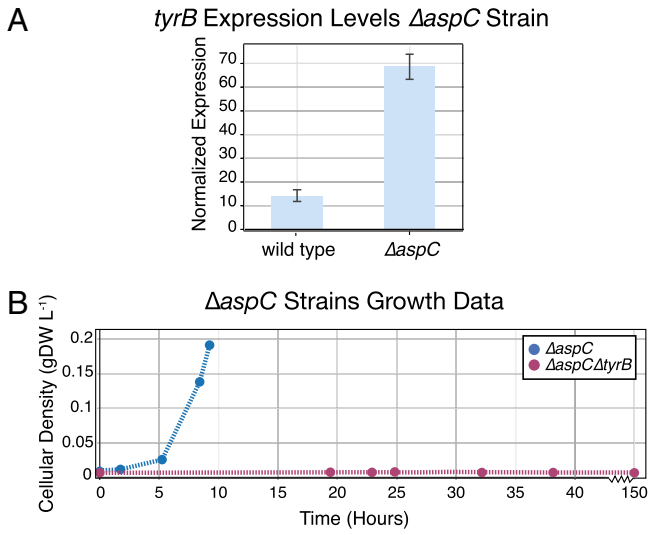

Fig. 2. Workflow-guided results used to discover isozymes of *aspC*. (*A*) Bar chart of the qPCR results in terms of normalized expression of the *tyrB* isozyme candidate in the Δ*aspC* and WT strains (SE ratio was calculated: $P < 0.05$; $n = 1$, two biological duplicates and six technical replicates). A fold increase of 4.8 is observed in the Δ*aspC* strain compared with WT. (*B*) Growth data on glucose minimal medium in terms of cellular density is reported for Δ*aspC* and Δ*aspC*Δ*tyrB* strains. The Δ*aspC*Δ*tyrB* strain shows no growth for >150 h. DW, dry weight.

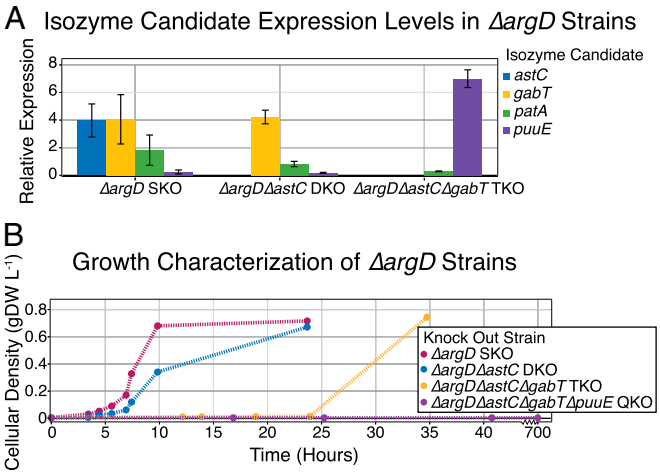

Fig. 3. Workflow-guided results used to discover isozymes for *argD*. (*A*) Bar chart of the relative expression, compared with WT, of candidate isozyme genes in the SKO, DKO, and TKO strains shows up-regulation that guided the multi-KO strain construction (SE ratio was calculated: $P < 0.05$; $n = 1$, two biological duplicates and six technical replicates). Note that *puuE* was not up-regulated until the construction of the Δ*argD*Δ*astC*Δ*gabT* TKO strain. (*B*) Growth data on glucose minimal medium in terms of cellular density are reported for the four strains iteratively constructed as guided by the workflow. The last strain constructed, Δ*argD*Δ*astC*Δ*gabT*Δ*puuE*, continued to show no growth after 700 h of incubation.

From the study of the *aspC* false-negative target, it was proposed that two functions should be considered to occur in *E. coli* K-12: (*i*) the annotated Tyr aminotransferase, *tyrB*, apart from being the "aromatic amino acid" aminotransferase, can also perform the role of Asp aminotransferase, and (*ii*) the Ile aminotransferase, *ilvE*, apart from being the "branched-chain amino acid" aminotransferase, can also perform the role of Tyr aminotransferase.

Case 2: *argD*–Acetylornithine Aminotransferase/*N*-Succinyldiaminopimelate Aminotransferase. Another member of a generalist enzymatic class (7) explored in this study was the enzyme encoded by the gene *argD*. This aminotransferase was previously identified as having dual functionality, involved in both Lys and Arg biosynthesis (26). The *argD* gene is predicted to be an essential gene because of its role in amino acid synthesis; however, KO studies have repeatedly shown the nonessentiality of this gene on glucose minimal medium (24). The putative isozyme targets explored for acetylornithine/*N*-succinyldiaminopimelate aminotransferase based on sequence homology were *astC*, *gabT*, *patA*, and *puuE*.

Initial examination of the growth of a Δ*argD* strain was performed. Following this confirmation of growth, RT-qPCR analysis was performed to examine the expression of isozyme candidate genes in the Δ*argD* strain compared with a WT strain (Fig. 3*A*). The candidate genes *astC* and *gabT* showed the greatest fold difference from WT in expression: 3.97-fold and 4.06-fold, respectively. The up-regulation of these two genes prompted the construction of two DKO strains, Δ*argD*Δ*astC* and Δ*argD*Δ*gabT*. Of the two, Δ*argD*Δ*astC* was the DKO strain initially chosen for examination due to a previously drawn relationship (27). The growth exhibited by this strain is displayed in Fig. 3*B*. The Δ*argD*Δ*astC* strain demonstrated only a mild difference in growth fitness compared with the Δ*argD* strain; therefore, further analysis of the remaining candidates was performed.

SYSTEMS BIOLOGY

A second round of RT-qPCR was performed on the Δ*argD*Δ*astC* DKO strain to identify isozyme candidates further. The gene *gabT* continued to be up-regulated in this DKO strain, with a relative expression ratio of 4.22 (Fig. 3*A*). This result led to the construction of a triple-knockout (TKO) strain, Δ*argD*Δ*astC*Δ*gabT*. A significant reduction in growth fitness was observed in the TKO strain, with an exhibited lag phase of ∼24 h (Fig. 3*B*). The eventual growth of the strain suggested the need for further examination of the remaining two candidates, *puuE* and *patA*.

A third round of RT-qPCR was performed on the constructed TKO strain. Although the *puuE* gene had been down-regulated in the SKO and DKO strains compared with the WT strain (relative expression ratios of 0.24 and 0.18, respectively), qPCR showed its up-regulation in the TKO strain (Fig. 3*A*). A relative expression ratio of 7.00 was found for the *puuE* gene, thereby prompting the construction of a quadruple-KO (QKO) strain. The strain Δ*argD*Δ*astC*Δ*gabT*Δ*puuE* was screened for growth for more than 4 wk in multiple attempts, and a conclusion of lethality was made. This result closed the experimental loop in the workflow. As a final validation, all remaining DKO and TKO combinations were constructed and their growth was validated to ensure the synthetic lethal interaction was as expected (Fig. S2*A*).

The results from examining isozymes of *argD* suggested the presence of a regulatory hierarchy regarding isozyme activation that emerged following serial KO and expression analysis of the multi-KO strains. The mechanisms influencing this regulatory response are suggested as an avenue for further study beyond the scope of work presented here.

To examine the possibility of acquired mutations in the strains constructed and used in this study, genome resequencing and analysis were performed for Δ*argD*, Δ*puuE*, and their descendent TKO strains. A summary of the results for these strains is provided in Table S3. The only obvious mutation of interest was that of the new junction call at *puuR*/*puuC* in the Δ*argD*Δ*astC*Δ*gabT* TKO strain. Further read-depth coverage analysis of this region revealed a 962-bp deletion between 1,360,264 and 1,361,226 bp, deleting a large section of the *puuR* gene (Fig. S2*B*). The genes *puuR* and *puuC* are both in the same operon as the isozyme *puuE* explored in this study, with *puuR* acting as a repressor for this operon under conditions of low putrescine concentration (28).

A potential mechanism for the large up-regulation of the *puuE* isozyme in the TKO strain is elucidated from a structural analysis of the repressor protein, *puuR*. The *puuR* DNA-binding transcriptional repressor consists of two domains, namely, a helix–turn–helix DNA-binding domain and a Cupin-family domain (28). Using the high number of homologous templates available, a homology model was constructed from the amino acid sequence of the *puuR* gene via the available toolkits (29–31) (with an average confidence of 97% across templates; Fig. S2*C*). The resulting structure demonstrates that the observed deletion in the Cupin domain from the read-depth coverage analysis is in direct contact with the helix–turn–helix motif in the DNA-binding domain. Thus, it was concluded that removing this part of the protein would drastically compromise protein integrity and prevent DNA binding, and, consequently, the ability of the *puuR* protein to repress the *puuE* gene.

Finally, examination of this case demonstrated the potential importance of extending incubation times in essentiality screens. Often, in high-throughput datasets, growth cutoff times are made for the sake of analysis (24, 32, 33), which could lead to misleading reports of essentiality. This study provided initial data on a range for incubation times required to make essentiality calls with higher accuracy. For the Δ*argD*Δ*astC*Δ*gabT* TKO strain, longer lag phases than those lag phrases typically observed in *E. coli* were measured. Interestingly, mutations were observed that were implicated in rescuing the growth and loss of the primary enzyme(s) under examination. As strain resequencing becomes more accessible, it is possible that similar mutations acquired during extended lag phases will be observed (34). As demonstrated here, strains exhibiting delayed or slow growth may present an interesting opportunity for discovery.

Case 3: *gltA*–Citrate Synthase. The last false-negative case examined in this study was the growth of a citrate synthase, *gltA*, KO strain. Previous studies demonstrated the ability of a 2-methylcitrate synthase to perform the same catalytic function as citrate synthase and suggested that mutagenesis is required for this transition to take place (35). Independently, using the approach presented here, a BLASTp analysis also pointed to *prpC* as a putative isozyme based on sequence homology. This case was thereby examined as a final validating case, demonstrating the up-regulation of an isozyme in the absence of the primary catalyzing enzyme.

Upon initial characterization of a Δ*gltA* strain, the strain did not grow on minimal medium despite the fact that it was listed as a positive growth phenotype in the initial Keio screen (24). Thus, the case could be considered a true-negative prediction; however, given the strong evidence for a possible homolog (Table S1), as well as previous literature reports (35), this case was further explored. Adaptive laboratory evolution was used to see if the putative isozyme, *prpC*, could indeed rescue a *gltA*-deficient strain when a selective pressure was applied. To promote growth, a "weaning" ALE was performed (36). FBA was conducted to determine which metabolite could be added to the medium to rescue growth (Table S4). It was predicted that supplementation with citrate would allow for utilization of glucose and support growth. Although other supplements were predicted to improve growth, citrate was selected due to its close relation to the citrate synthase reaction, as well as the inability of *E. coli* to use citrate as its sole carbon source, thereby forcing metabolism of glucose (37). The Δ*gltA* was therefore grown with citrate supplementation in two parallel ALE experiments. Growth was observed in Δ*gltA* ALE experiment 1 (ALE 1) and Δ*gltA* ALE 2 with supplementation, and following two passages, robust growth was observed without supplementation (Fig. 4*A*). The final apparent cell densities for each ALE experiment showed approximate eightfold and sixfold increases (ALE 1 and ALE 2, respectively) from the initial supplemented state.

Expression analysis of the evolved end points showed significant up-regulation (more than 500-fold difference) of the isozyme target, *prpC*, thereby providing evidence of its isozyme function and allowing for a linear progression through the workflow (Fig. 4*B*). Furthermore, the growth rate of each isolate clone correlated well with the expression level of *prpC* (Fig. S3*D*). Following observation of these results, an attempt was made to knock out *prpC* from the four end-point clones. Although KO confirmation primers suggested successful removal of the *prpC* gene, the strains continued to demonstrate growth. This positive growth result suggested the possibility of duplication of the *prpC* gene elsewhere in the genome. Therefore, in-depth, whole-genome resequencing and analysis were performed.

Two clones from each experiment, Δ*gltA* from ALE 1 and Δ*gltA* from ALE 2, were isolated and resequenced along with the parent Δ*gltA* Keio strain. Resequencing results revealed the possibility of new junction evidence and mutations within the *prp* operon (Table S3) for three of the four end-point isolates. Further analysis showed elevated read coverage of the region between 279 and 371 kb in the ALE end-point clones but not in the Δ*gltA* parent strain (Fig. S3*B*). The same coverage abnormality was also confirmed in ALE clones isolated from the third flask of the ALE experiments (the experiments consisted of 8 flasks total) (Fig. S3*A*). A single relevant novel junction was then identified by applying a custom pipeline based on de novo fragment assembly, providing significant evidence for tandem duplication. The presence of the junction was further verified in all end-point isolates by performing read mapping onto the

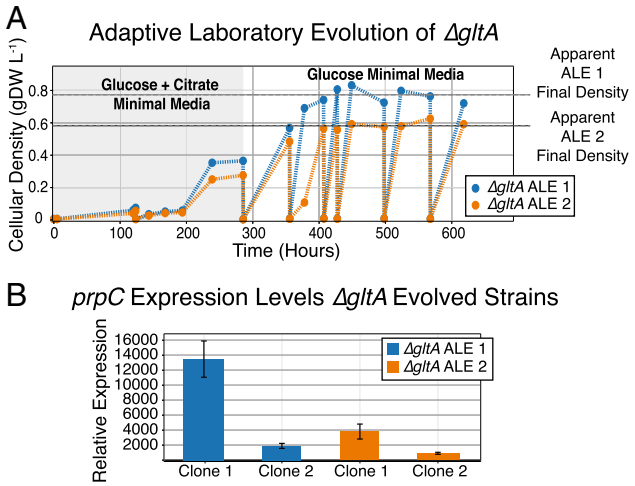

Fig. 4. Workflow-guided results used to discover isozymes of *gltA*. (*A*) Cellular density results from the ALE of Δ*gltA* on glucose minimal medium are illustrated. A vertical drop in cellular density corresponds to manual passaging of a fraction of the cell culture for a fresh batch of medium. The independent ALE experiments reached a different apparent final density. (*B*) Bar chart shows qPCR results as a fold increase in expression of the *prpC* isozyme candidate in four ALE end-point clones in relation to a WT strain (SE ratio was calculated: $P < 0.05$; $n = 1$, two biological duplicates and six technical replicates).

corresponding sequence. The duplicated region was flanked by a 182-bp repeat, which is a part of the IS30 mobile element. The rightmost flanking copy is not present in the GenBank reference K-12 strain (38) and is unique to the Keio K-12 parent strain. Furthermore, plots from Fig. S3 clearly suggest high multiplicity of the 100-kb duplicated fragment. To perform the copy number analysis, per base coverage data were normalized to the average coverage of the particular strain and to the position-specific biases, inferred from the Keio parent strain coverage distribution (Fig. S3*C*). Then, the multiplicity was estimated by the median of the normalized coverage values across the duplicated region. Predicted multiplicities of the duplicated region are 7, 11, 16, and 8 for the different clones, respectively. Finally, analysis of novel junctions also predicted a smaller scale, 1-kb duplication event in ALE 1 clone 1. The genome coordinates of this duplication are 348,810–349,895, spanning part of the *prpC* gene and encompassed by the larger scale 100-kb duplication as well.

The results from the coverage and whole-genome resequencing analysis helped to explain the inability to construct the DKO strain. Based on the suspected high copy number of duplication, it was deemed unreasonable to knock out all copies of the *prpC* gene, and the experimental workflow was concluded. Thus, given the transcriptional and mutational evidence, it is likely that the *prpC* gene is indeed an isozyme for *gltA*, as has previously been reported. Furthermore, these results expand upon the theories surrounding previous reports of genome duplication amplifications as an evolutionary mechanism (11, 34).

Insight into biological adaptability requiring evolution was thus gained from exploring the presence of large-scale duplications. The mutation event was required to rescue growth and activate the known isozyme, *prpC*, similar to a previous study (35). This mutation event occurred after two ALE passages in this study (Fig. S3). Interestingly, all four individual end-point clones, as well as the two clones from pass 3 that were isolated and sequenced, exhibited the same large-scale duplication of the 100-kb region, thereby implying a clear evolutionary pressure to up-regulate this particular region. Published theories regarding genome duplication amplifications have remarked on the instability of large duplications and their subsequent loss in the absence of selection pressures (11). Thus, although the duplications were detected after the third passage and, again, after the eighth passage, they could be lost with further adaptation.

Gene–Protein-Reaction Analysis and Conservation of Isozymes Across 55 Related Strains. Inconsistencies between in silico predictions and in vivo data guided this study and resulted in the discovery of seven links between known, partially characterized enzymes and reactions that are conditionally essential to the metabolic network in *E. coli*. Suggested changes to the gene–protein-reaction (GPR) associations in *i*JO1366 based on genetic and transcriptional evidence are presented in Fig. 5 (reaction abbreviations are provided in Table S5). The expanded reaction associations for the cases examined in this study support the hypothesis that functional overlap occurs for enzymes across metabolism, forming the basis of an underground metabolic network, and this concept has been supported by other recent works (6, 8).

As a preliminary expansion of this study, the conservation of the newly discovered isozymes in 55 closely related strains of *E. coli* and *Shigella* that have existing metabolic models was investigated (39). It was determined that the same GPR changes should be made in the majority of these models. However, some of the putative isozymes discovered in this study have no corresponding gene in the related strains (Fig. S4). For example, eight *Shigella* strains examined are lacking *prpC*, the *gltA* isozyme. Also, 14 *E. coli* strains from different clades lack *puuE*, one of newly discovered *argD* isozymes. Finally, five of the *Shigella* strains lack *astC*, another one of the *argD* isozymes. Therefore, new GPR associations are available for each of the 55 models, but they must be adjusted in a strain-specific manner. Furthermore, analysis of isozyme and regulatory region sequence conservation between different strains of *E. coli* could illuminate divergent evolutionary strategies in the *E. coli* species.

Conclusions

Enzymatic promiscuity and a cell's ability to adapt to genetic perturbations were explored in the execution of the model-driven workflow developed in this study. The results suggest that a hierarchy of latent metabolic solutions exists, as highlighted by the analysis of the false-negative *argD*. Furthermore, this study emphasized the possibility of discovering novel regulatory responses following long-term culturing studies and genome resequencing when determining gene essentiality. Lastly, the developed methods can be readily extended to other organisms and gene targets where gap-filling is required. For example, a gap-filling study using the *E. coli* *i*JO1366 metabolic

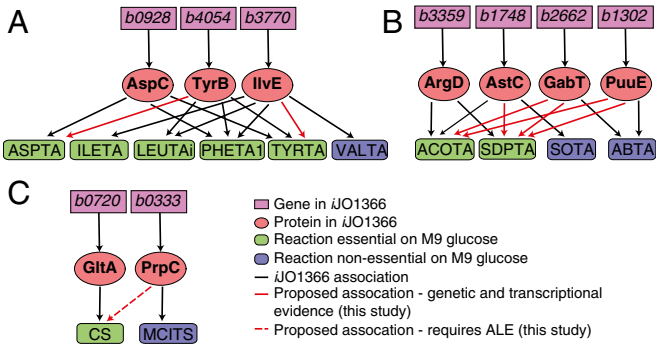

Fig. 5. Summary of GPR associations to be added to the *E. coli* metabolic network reconstruction *i*JO1366 based on findings from three cases: *gltA*, *aspC*, and *argD*. Proposed associations are highlighted in red. (*A*) GPR additions for *tyrB* and *ilvE*. (*B*) GPR additions for the quadruple synthetic lethal interaction set. (*C*) GPR for *gltA* and *prpC* highlights the requirement of evolution or mutagenesis for the suggested association. Reaction abbreviations are from the iJO1366 reconstruction.

SYSTEMS BIOLOGY

reconstruction has identified a total of 265 false-negative predictions (corresponding to 59 unique genes), which could be explored under various environmental and genetic conditions (14). The extendability of this work to pathogenic organisms could be particularly advantageous in searching for antimicrobial targets. As genome-scale models and organism-specific knowledge bases expand, their ability to predict biological behavior for both basic science and biotechnology applications will increase. This likely expansion is evident in the appearance of complementary studies (8) that use computational modeling to isolate specific predicted functionalities through gene KO or overexpression and through determining media conditions that focus pressure on the predicted function(s). A comparison of the present study with the aforementioned study (8) shows some overlap between genes explored and thought to exhibit underground activity (*ilvE* and *tyrB*), although the suggested activities reported in each study for these genes are distinct. Workflows for discovering promiscuous and latent activities, such as the one presented here, will be critical for advancement of model-driven science.

Although the strengths of the presented method were demonstrated with the cases explored in this study, there is room for improvement to broaden the applicability of the workflow. For false-negative model gaps, there is the possibility that an alternate pathway is rescuing the growth of the cell. For such alternate pathway solutions, isozyme analysis could result in fruitless effort because those solutions are not captured by the workflow. Another area of improvement proposed for the workflow is in selecting bioinformatic algorithms. There are many enzymes for which BLASTp, a purely sequence homology-driven algorithm, will not result in the identification of candidate isozymes. To expand this list of putative isozymes, the use of protein structure similarity or substrate structure similarity identifying algorithms is suggested (40). Finally, the utilization of RNA-sequencing or other larger scale "omics" methods could capture a more complete picture of transcriptional changes in response to KO perturbations rather than qPCR analysis. These modifications to the workflow presented may result in a more robust method for filling model gaps that can be applied to other organisms as well.

Materials and Methods

The isozyme discovery workflow presented in this study was executed utilizing materials and methods that are described in detail in *SI Materials and Methods*. FBA simulations and bioinformatics tools used to identify false negative targets and corresponding isozyme candidates as well as predict nutrient supplementation are described in *SI Materials and Methods*. The qPCR protocol utilized to analyze up-regulation of candidate isozymes in strains examined is described in *SI Materials and Methods*. KO strains constructed in this study and their associated parent Keio (24) strains are summarized in Table S6. Culture conditions and growth characterizations on M9 minimal medium are described in *SI Materials and Methods*. Furthermore, details regarding the ALE by weaning off of supplementation are provided in *SI Materials and Methods*. Genome resequencing was performed to assess the accumulation of mutations in the strains characterized on glucose minimal medium in comparison to the parent strains grown on rich medium. The details related to the construction of paired-end resequencing libraries as well as associated mutation analysis is described in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Zachary King for helpful discussions and insight, Richard Szubin for assistance with strain resequencing, and Daniel Zielinski for inspiration. This work was partially supported by the Novo Nordisk Foundation and NIH Grants/National Institute of General Medical Sciences Grant 1R01GM057089.

1. Tawfik DS (2014) Accuracy-rate tradeoffs: How do enzymes meet demands of selectivity and catalytic efficiency? *Curr Opin Chem Biol* 21:73–80.
2. Jensen RA (1976) Enzyme recruitment in evolution of new function. *Annu Rev Microbiol* 30:409–425.
3. Lazcano A, Miller SL (1996) The origin and early evolution of life: Prebiotic chemistry, the pre-RNA world, and time. *Cell* 85(6):793–798.
4. Rison SCG, Thornton JM (2002) Pathway evolution, structurally speaking. *Curr Opin Struct Biol* 12(3):374–382.
5. Khersonsky O, Tawfik DS (2010) Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annu Rev Biochem* 79:471–505.
6. D'Ari R, Casadesús J (1998) Underground metabolism. *BioEssays* 20(2):181–186.
7. Nam H, et al. (2012) Network context and selection in the evolution to enzyme specificity. *Science* 337(6098):1101–1104.
8. Notebaart RA, et al. (2014) Network-level architecture and the evolutionary potential of underground metabolism. *Proc Natl Acad Sci USA* 111(32):11762–11767.
9. Furnham N, de Beer TAP, Thornton JM (2012) Current challenges in genome annotation through structural biology and bioinformatics. *Curr Opin Struct Biol* 22(5): 594–601.
10. van Loo B, et al. (2010) An efficient, multiply promiscuous hydrolase in the alkaline phosphatase superfamily. *Proc Natl Acad Sci USA* 107(7):2740–2745.
11. Andersson DI, Hughes D (2009) Gene amplification and adaptive evolution in bacteria. *Annu Rev Genet* 43:167–195.
12. McCloskey D, Palsson BO, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol Syst Biol* 9:661.
13. Bordbar A, Monk JM, King ZA, Palsson BO (2014) Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet* 15(2):107–120.
14. Orth JD, Palsson B (2012) Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions. *BMC Syst Biol* 6:30.
15. Reed JL, et al. (2006) Systems approach to refining genome annotation. *Proc Natl Acad Sci USA* 103(46):17480–17484.
16. Kumar VS, Maranas CD (2009) GrowMatch: An automated method for reconciling in silico/in vivo growth predictions. *PLOS Comput Biol* 5(3):e1000308.
17. Orth JD, et al. (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. *Mol Syst Biol* 7:535.
18. Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. *Curr Opin Biotechnol* 14(5):491–496.
19. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
20. Fotheringham IG, et al. (1986) The cloning and sequence analysis of the aspC and tyrB genes from Escherichia coli K12. Comparison of the primary structures of the aspartate aminotransferase and aromatic aminotransferase of E. coli with those of the pig aspartate aminotransferase isoenzymes. *Biochem J* 234(3):593–604.
21. Gelfand DH, Steinberg RA (1977) Escherichia coli mutants deficient in the aspartate and aromatic amino acid aminotransferases. *J Bacteriol* 130(1):429–440.
22. Powell JT, Morrison JF (1978) Role of the Escherichia coli aromatic amino acid aminotransferase in leucine biosynthesis. *J Bacteriol* 136(1):1–4.
23. Lee-Peng FC, Hermodson MA, Kohlhaw GB (1979) Transaminase B from Escherichia coli: Quaternary structure, amino-terminal sequence, substrate specificity, and absence of a separate valine-alpha-ketoglutarate activity. *J Bacteriol* 139(2):339–345.
24. Baba T, et al. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. *Mol Syst Biol* 2:0008.
25. Barrick JE, et al. (2009) Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* 461(7268):1243–1247.
26. Ledwidge R, Blanchard JS (1999) The dual biosynthetic capability of N-acetylornithine aminotransferase in arginine and lysine biosynthesis. *Biochemistry* 38(10):3019–3024.
27. Newman J, et al. (2013) Determination of the structure of the catabolic N-succinylornithine transaminase (AstC) from Escherichia coli. *PLoS ONE* 8(3):e58298.
28. Nemoto N, et al. (2012) Mechanism for regulation of the putrescine utilization pathway by the transcription factor PuuR in Escherichia coli K-12. *J Bacteriol* 194(13): 3437–3447.
29. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: A unified platform for automated protein structure and function prediction. *Nat Protoc* 5(4):725–738.
30. Hildebrand A, Remmert M, Biegert A, Söding J (2009) Fast and accurate automatic structure prediction with HHpred. *Proteins* 77(Suppl 9):128–132.
31. Kelley LA, Sternberg MJ (2009) Protein structure prediction on the Web: A case study using the Phyre server. *Nat Protoc* 4(3):363–371.
32. Nichols RJ, et al. (2011) Phenotypic landscape of a bacterial cell. *Cell* 144(1):143–156.
33. Gerdes SY, et al. (2003) Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *J Bacteriol* 185(19):5673–5684.
34. Finkel SE (2006) Long-term survival during stationary phase: Evolution and the GASP phenotype. *Nat Rev Microbiol* 4(2):113–120.
35. Patton AJ, Hough DW, Towner P, Danson MJ (1993) Does Escherichia coli possess a second citrate synthase gene? *Eur J Biochem* 214(1):75–81.
36. Lee DH, Palsson BO (2010) Adaptive evolution of Escherichia coli K-12 MG1655 during growth on a Nonnative carbon source, L-1,2-propanediol. *Appl Environ Microbiol* 76(13):4158–4168.
37. Koser SA (1924) Correlation of citrate utilization by members of the colon-aerogenes group with other differential characteristics and with habitat. *J Bacteriol* 9(1):59–77.
38. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2005) GenBank. *Nucleic Acids Res* 33(Database issue):D34–D38.
39. Monk JM, et al. (2013) Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proc Natl Acad Sci USA* 110(50):20338–20343.
40. Zhang QC, et al. (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* 490(7421):556–560.