PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Systems-Wide Prediction of Enzyme Promiscuity Reveals a New Underground Alternative Route for Pyridoxal 5'-Phosphate Production in *E. coli*

Matthew A. Oberhardt[1,2,3☯]*, Raphy Zarecki[1☯], Leah Reshef[2], Fangfang Xia[4], Miquel Duran-Frigola[5], Rachel Schreiber[2], Christopher S. Henry[4], Nir Ben-Tal[6], Daniel J. Dwyer[7‡], Uri Gophna[2‡], Eytan Ruppin[1,3‡]*

1 School of Computer Sciences and Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel, **2** Department of Molecular Microbiology and Biotechnology, Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel, **3** Center for Bioinformatics and Computational Biology, Department of Computer Science, University of Maryland, College Park, Maryland, United States of America, **4** Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, United States of America, **5** Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain, **6** Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel, **7** Department of Cell Biology and Molecular Genetics, Institute for Physical Science and Technology, Department of Bioengineering, Maryland Pathogen Research Institute, University of Maryland, College Park, Maryland, United States of America

☯ These authors contributed equally to this work.
‡ DD, UG, and ER are equally contributing last authors.
* mattoby@gmail.com (MAO); ruppin@post.tau.ac.il (ER)

Citation: Oberhardt MA, Zarecki R, Reshef L, Xia F, Duran-Frigola M, Schreiber R, et al. (2016) Systems-Wide Prediction of Enzyme Promiscuity Reveals a New Underground Alternative Route for Pyridoxal 5'-Phosphate Production in *E. coli*. PLoS Comput Biol 12(1): e1004705. doi:10.1371/journal.pcbi.1004705

Editor: Bernhard O. Palsson, University of California San Diego, UNITED STATES

Received: June 17, 2015

Accepted: December 14, 2015

Published: January 28, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Funding agencies: (MO) Whitaker Foundation (Whitaker International Scholars Program) (http://www.whitaker.org/grants/fellows-scholars) (MO) Dan David Fellowship (http://www.dandavidprize.org/scholarship-applications) (ER) European Union FP7 INFECT project (http://www.fp7infect.eu/) ERA-Net Plant project (http://www.erapg.org/publicpage.m?key=everyone&trail=/everyone)

Abstract

Recent insights suggest that non-specific and/or promiscuous enzymes are common and active across life. Understanding the role of such enzymes is an important open question in biology. Here we develop a genome-wide method, PROPER, that uses a permissive PSI-BLAST approach to predict promiscuous activities of metabolic genes. Enzyme promiscuity is typically studied experimentally using multicopy suppression, in which over-expression of a promiscuous 'replacer' gene rescues lethality caused by inactivation of a 'target' gene. We use PROPER to predict multicopy suppression in *Escherichia coli*, achieving highly significant overlap with published cases (hypergeometric p = 4.4e-13). We then validate three novel predicted target-replacer gene pairs in new multicopy suppression experiments. We next go beyond PROPER and develop a network-based approach, GEM-PROPER, that integrates PROPER with genome-scale metabolic modeling to predict promiscuous replacements via alternative metabolic pathways. GEM-PROPER predicts a new indirect replacer (*thiG*) for an essential enzyme (*pdxB*) in production of pyridoxal 5'-phosphate (the active form of Vitamin B_6), which we validate experimentally via multicopy suppression. We perform a structural analysis of *thiG* to determine its potential promiscuous active site, which we validate experimentally by inactivating the pertaining residues and showing a loss of replacer activity. Thus, this study is a successful example where a computational

(ER) I-CORE Program of the Planning and Budgeting Committee and The Israel Science Foundation (grant No 41/11) (www.i-core.org.il/ISF) (UG) McDonnell foundation (https://www.jsmf.org/) (UG) German-Israeli Project Cooperation (DIP) (http://www.dfg.de/en/research_funding/programmes/international_cooperation/german_israeli_cooperation/) (MD) Spanish FPU grant (http://cepima.upc.edu/positions/FPU_2013) (MD) FEBS short term fellowship (http://www.febs.org/our-activities/fellowships/febs-short-term-fellowships/guidelines-for-febs-short-term-fellowships) (NBT) Grant No. 1775/12 of the I-CORE Program of the Planning and Budgeting Committee and The Israel Science Foundation

Competing Interests: The authors have declared that no competing interests exist.

investigation leads to a network-based identification of an indirect promiscuous replacement of a key metabolic enzyme, which would have been extremely difficult to identify directly.

Author Summary

Many enzymes can perform secondary functions at low affinities or rates, but such 'promiscuous' functions have never been predicted on a genome-wide scale. Here, we present the first genome-wide method to predict promiscuous functions of metabolic genes, which we apply to *E. coli*. Notably, we predict and validate several new cases where a 'replacer' gene can compensate for the loss of an essential 'target' gene through a promiscuous activity. Next, we couple our method with a genome-scale metabolic model, in order to search for 'replacer' genes that compensate for essential 'target' genes by metabolically bypassing them. We use this network-augmented approach to uncover a novel promiscuous pathway for the production of pyridoxal 5'-phosphate (the active form of Vitamin B_6) in *E. coli*. This study represents an important step in understanding promiscuous functions in bacteria, and is a prime example of a systems-level analysis leading to new biological insight.

Introduction

Enzymes have traditionally been associated with discrete activities [1]. Clear-cut annotations populate well-known enzyme databases such as KEGG and Uniprot, and foster an implicit assumption that gene activities are specific. However, recent advancements in our understanding of evolution and enzyme activity have cast this view into question, and suggest instead that non-optimized and/or promiscuously active enzymes are frequent and active across life [2]. 'Generalist' enzymes (i.e., those carrying more than one function) have been shown to be abundant, to play different biological roles than 'specialist' enzymes, and to behave differently than specialist enzymes during switches in media conditions [3]. Promiscuous enzyme activities (i.e., those that are only active with low affinities/activities) have been shown to play adaptive biological roles (e.g., [4] and [5]), and to often arise through neutral mutations that are not detrimental to the primary enzymatic activity [6]. In particular, antibiotic resistance may emerge initially through the action of over-expressed, promiscuous genes [7, 8].

Because of recent conceptual and modeling advances, there is significant interest in being able to predict and exploit enzyme promiscuity [9–12]. Tools have been developed that mine molecular signatures or three-dimensional catalytic domains of enzymes to predict whether enzymes are promiscuous [13–15], and these methods have been used, e.g., to design likely retrosynthetic pathways [9, 11]. While these methods achieve important goals, they have not been aimed at evaluating the functional implications of *existing* promiscuous activities on a genome-wide, network scale. A notable recent effort did probe the global relevance of promiscuous enzyme activities, but the study used 'as is' data collected from enzyme databases, which, while having the advantage of being fully experimentally verified, include many non-biologically-relevant instances and exclude many biologically relevant ones [16]. Another recent study used knowledge gaps identified in a genome-scale metabolic model of *E. coli* to identify candidate isozymes, which display substrate promiscuity [17]. In that study, a 'target' enzyme that is found to be nonessential *in vitro* despite being essential *in silico* is knocked out, after which potential isozymes that are found to be upregulated are cumulatively knocked out until the cell

can no longer survive. This method thus identifies enzymes with a secondary isozyme function that, when the cell overexpresses them, can compensate for the loss of the 'target'. Interestingly, in one case they consider, the target is in fact essential, and adaptive evolution is required for the potential isozyme to be expressed enough to compensate for its loss. Searching for many more such low activity promiscuous functions in an unsupervised, genome-wide way is the focus of this present study.

To accomplish this, we utilize an unsupervised PSI-BLAST based method for assessing potential secondary functions of genes in *E. coli*. We predict promiscuous 'replacer' functions that may compensate for primary 'target' functions in other genes if they are altered or lost, and compare these predictions to known cases in which an over-expressed gene can take on a secondary role, as shown in an assay generally referred to as multicopy suppression [18]. Multicopy suppression reveals 'replacer' genes whose over-expression suppresses the lethality caused by knocking out conditionally essential 'target' genes. It is thus an elegant assay for uncovering promiscuous gene functions. We find that our method for predicting promiscuous functions successfully recapitulates known multicopy suppression events [8, 19] and predicts new ones, several of which we validate *in vitro*.

We next develop a genome-scale metabolic modeling (**GEM**) based approach to predict and experimentally test cases of multicopy suppression in which functional promiscuous replacement happens 'indirectly' through a bypassing pathway or reaction, rather than by directly replacing the function of the target gene. For this, we focus on the gene target *pdxB*, for which multiple promiscuous replacers have been reported in the past [8]. PdxB is a key enzyme for producing pyridoxal 5'-phosphate (the active form of Vitamin B6), an essential nutrient in *E. coli*. Our experiment proceeded in 4 steps: First, we predicted that the gene *thiG* harbors a promiscuous activity that metabolically bypasses *pdxB*. Second, we validated this prediction with a new *in vitro* individual multicopy suppression experiment. Third, we performed a detailed structural analysis of ThiG and hypothesized which residues perform the promiscuous function. And Fourth, we mutated this active site and confirmed loss of *pdxB* replacement activity. Thus, we begin at the level of a genome-wide screen for promiscuous functions, and proceed to identify and validate a promiscuous pathway for production of an essential nutrient in *E. coli*, which could have important functional implications for this model organism.

Results

Developing an enzyme PROmiscuity PrEdictoR (PROPER)

We aimed to predict promiscuous functions of genes in a systematic and unsupervised manner, and specifically to do it on a genome-wide scale. We started by doing permissive PSI-BLAST based searches for distant gene similarities across RAST, a large consistent database of metabolic reactions and compounds that includes thousands of microbes [20]. This search allowed us to assign putative secondary functions for each initial "root" gene in the search (see Methods and Fig 1). We did this across metabolic genes in *E. coli*, which resulted in a set of phylogenetic trees, one for each root gene, that include up to thousands of genes from other bacteria along with their associated functions.

Using these trees, we searched for instances in which a gene in one of the trees had an assigned function different from the one that the RAST database assigns to the root gene. We restricted our search to genes whose primary functions are metabolic, as this is our focus of interest. Since our tree building method was more permissive than typical BLAST or PSI-BLAST, we considered these predicted functions as potential secondary functionalities that might only be effective at high expression levels or with slight mutations. We entered these cases in a large gene-by-function matrix.

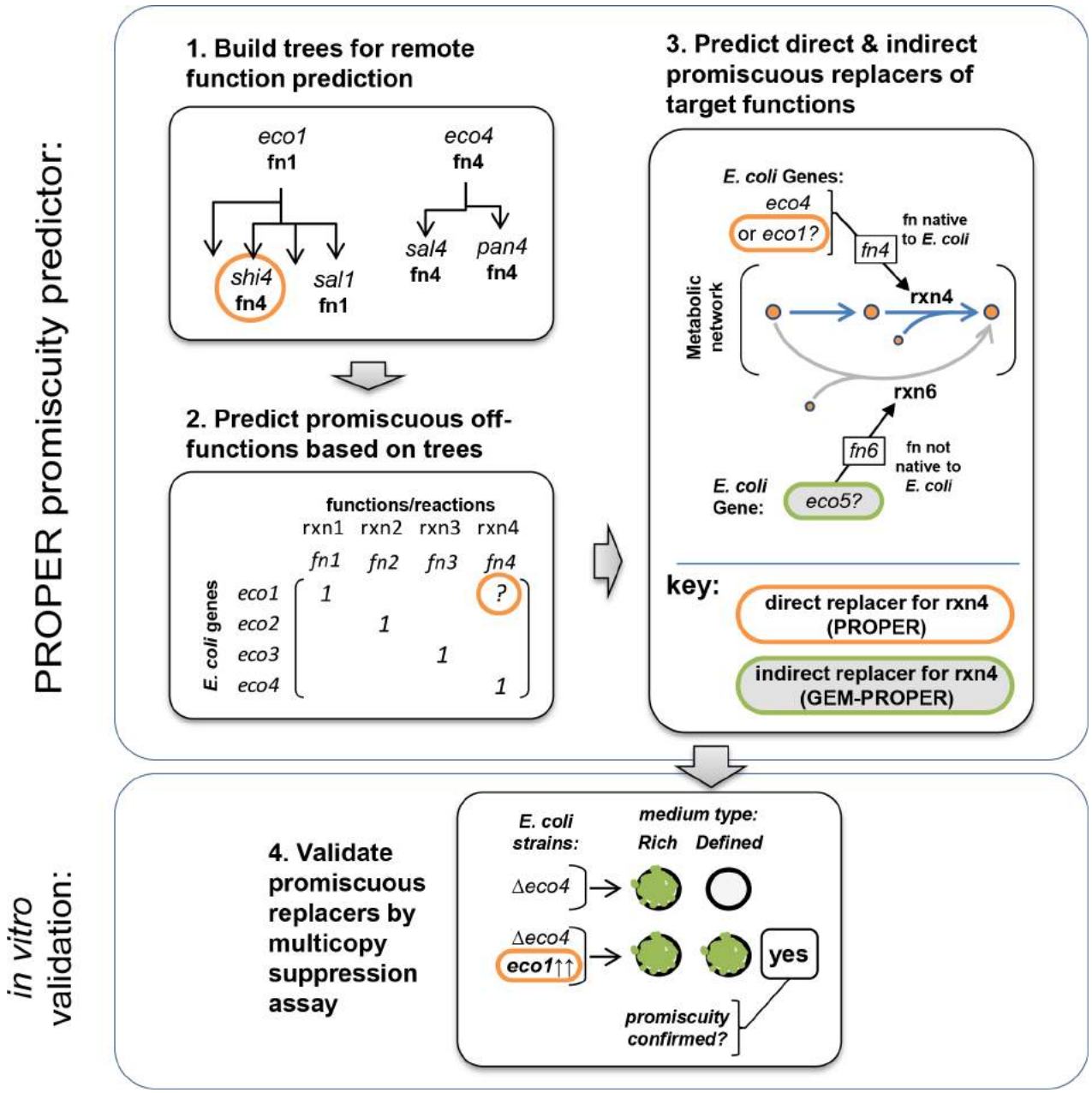

Fig 1. Schematic of prediction framework for promiscuous replacers. (1) Gene similarity trees are built around each gene in *E. coli*, including any distantly related gene in the RAST database. (2) A matrix is formed which links genes with their primary functions and also potential promiscuous functions. A gene (in this example, *eco1*) will take a potential secondary 'promiscuous' function in the matrix if its similarity tree includes any genes annotated with different functions (e.g., in this example, *shi4*, which encodes function fn4). (3) Cases in which a gene's predicted promiscuous function is identical to the function of another gene in *E. coli* constitute predicted 'direct' target-replacer gene pairs (via PROPER). We also predict 'indirect' target-replacer pairs where a replacer bypasses the target's function (via GEM-PROPER). (4) Promiscuous activity of a 'replacer' gene can be confirmed for target-replacer pairs in which the target is conditionally essential on a minimal medium, via the multicopy suppression assay.

doi:10.1371/journal.pcbi.1004705.g001

To facilitate testing and validation, we used our promiscuity matrix to predict pairs of genes that correspond to those identified through an *in vitro* assay called Multicopy suppression [19]. Multicopy suppression tests 'replacer' genes for promiscuous activities that either directly (catalyze similar reactions) or indirectly (catalyze distinct reactions) replace the function of a

'target' gene and preserve growth. It proceeds in four steps: (1) a 'target' gene is identified in the bacterium of interest (target genes must be essential on a certain test medium [usually M9], but not on a rich medium); (2) a strain is created in which the target gene is inactivated (due to the condition in step 1, this strain is not viable on the test medium); (3) a library of native genes (hereafter called 'replacers') is individually over-expressed in the organism on high-copy plasmids; and (4) transformed strains are grown on the test medium, and survivors are noted as having identified successful target-replacer gene pairs. Importantly, the replacer genes are native to the genome of the organism, but their natural expression dosage is not sufficient to promiscuously support growth. Hence, replacer genes discovered through multicopy suppression are hypothesized to possess secondary, low-affinity/low-activity functions that compensate for the function of the target. We predict target-replacer pairs in two ways: direct replacements, and indirect replacements.

To predict direct replacements, we search our matrix for instances in which one 'replacer' gene in *E. coli* has a secondary function assigned through our method, which corresponds exactly to the function of some other 'target' gene that is also found in *E. coli*. We call this straightforward method the enzyme PROmiscuity PrEdictoR, or PROPER.

To predict indirect replacements, we use the added insight of a GEM of *E. coli* (obtained from SEED: [21]). The SEED *E. coli* GEM was used, as opposed to one of the highly curated *E. coli* GEMS such as [22], because of our requirement that metabolic reactions from the SEED database can be seamlessly added to the GEM to test their effect on cellular fluxes (which would first require a large reconciliation of nomenclature if using a non-SEED GEM). Specifically, we search for replacers that, through a promiscuous function identified in our matrix, metabolically bypass the need for the target. These predictions were done in three steps: (1) we knocked out the target gene in a GEM (this could only be done for genes that were conditionally lethal on M9 *in silico*); (2) we searched in our gene similarity trees for any *E. coli* genes that harbor non-*E. coli* promiscuous functions; (3) we added each of these promiscuous functions in turn to the GEM, and looked for any that rescued *in silico* growth. This GEM-based PROmiscuity PrEdictoR (GEM-PROPER) thus identified indirect replacer genes that act through metabolism (see panel 3 of Fig 1).

In all, PROPER predicted 2811 direct target-replacer pairs in *E. coli*, encompassing 794 metabolic target genes and 753 metabolic replacer genes (see S1 Table). GEM-PROPER predicted a total of 98 indirect target-replacer pairs in *E. coli* (see full list in S4 Table). Since direct replacers span a large portion of metabolism, we focus on these predictions for our large-scale analyses (i.e., the first several sections of the results). Afterwards, we shift our focus to GEM-PROPER, and present a detailed analysis of a novel indirect predictor that we validated.

Systems-wide testing of PROPER

The number of direct replacers per target follows an exponentially decaying distribution (see S1 Fig), implying that promiscuous gene functions cluster around key target functions. The target genes with the most putative replacers are involved in fatty acid degradation or synthesis [*fadD* and *fadK* (degradation); *fabD* (synthesis)], or transport processes [*malF* (maltose/maltodextrin transport); *potC*, *ydcV*, *potB*, and *ydcU* (spermidine/putrescine transport); and *cysW* (sulfate transport)]. Fatty acid degradation and synthesis processes are known to involve enzymes with many potential specificities depending on only small alterations [23].

To validate our direct predictions, we collected results from two studies of multicopy suppression in *E. coli*, which between them had discovered 48 instances of multicopy suppression among 21 target genes [8, 19]. Both of these assays looked for multicopy suppressors of targets that were essential on M9 medium but not in a rich medium, and one used a combinatorial

assay that assessed every gene in *E. coli* as a potential replacer [19]. All but 3 of the target genes for which replacers had been found were metabolic, supporting our focus on metabolism. In all, we predicted a total of 63 replacers for the 21 targets, versus the 48 found previously by multicopy suppression. The replacers we predicted overlapped with those found previously by 8 genes. Notably, for each of the 21 target genes, there were over a thousand metabolic genes that could serve as potential replacers; therefore, the chance of achieving this level of overlap between our predictions and those found experimentally through random guesses is extremely low (p = 4.5e-15 in a hypergeometric test).

As another control test, we performed a BLAST search of each target gene identified in Patrick *et. al.* against all metabolic genes in *E. coli*, and kept only those above the default threshold as potential replacers. This yielded 122 predicted replacers for the 21 target genes, among which none matched results from Patrick *et. al.*, and only two overlapped with predictions made by PROPER (neither of those turned out true in our own experiments). This emphasizes the added value of PROPER over simple BLAST searches for finding find such low affinity promiscuous functions.

We also considered it possible that some of our target-replacer predictions are true, but cannot be verified using multicopy suppression because the target gene is not conditionally lethal (which is a necessary condition for doing a multicopy suppression experiment–see Fig 1, panel 4). As a quick test of this, we compared our direct predictions to the isozyme sets *aspC/tyrB*, *argD/astC/gabT/puuE*, and *gltA/prpC*, which were discovered in [17] as described in the introduction. Indeed, all 8 pairwise combinations of genes from within any of these 3 isozyme sets comes up as a *reciprocal* target-replacer pair in our direct predictions, meaning that either gene in each pair can play the role of the target or the replacer. That all of these pairs would come up in our predictions and, given that, that they would all be reciprocal are both highly unlikely by chance (p = 3e-19 and p = 3e-5 in hypergeometric tests, denoting that all pairs from [17] (a) overlap our predictions vs. a background of all-vs.-all pairings of our predicted targets and replacers, and (b) are reciprocal, assuming the same likelihood for any pair). This suggests that promiscuous functions that have enough activity to be considered isozymes (or, more specifically, to be defined so through the experimental pipeline in [17]) may tend to come up in our predictions as reciprocal, and suggests a complimentary association between our methods and those in the aforementioned work.

Testing novel predicted direct replacers via new multicopy suppression assays

The random gene insertion method used by Patrick et al. to screen for potential target-replacer pairs [19] is demonstrably not comprehensive (e.g., 7 replacers were found for *pdxB* in [8]; none of these were reported in the Patrick study for *pdxB*, although *pdxB* was in the set of replacers that Patrick et al. tested). Hence, we next asked how many of our novel predicted target-replacer pairs (those that had not been previously found experimentally) may be true. To test this we developed a single-replacer multicopy suppression assay analogous to that used to validate individual results in the initial multicopy suppression study by Patrick [19]. This assay involves plating target-replacer strains (i.e., strains with the target knocked out, and the replacer over-expressed on a plasmid) on M9 medium, and observing how long it takes for colonies to appear. We judged successful target-replacer pairs as those that grew more robustly than an empty plasmid control (see Methods, as well as Supplementary methods in S1 Text for a fuller explanation of this assay).

Using these criteria, we verified 7 of the 8 target-replacer pairs that were identified in Patrick *et. al.* and predicted by PROPER (see S2 Table for full list of experimental results). We then

assessed each of our 55 novel predicted target-replacer pairs for multicopy suppression, and found three of them to have replacer activity (*hisA*, *cysM*, and *metB* were viable replacers for *hisH*, *ilvA*, and *metC* respectively; see S2 Table). Thus, in all we validated 10 out of 63 direct replacement predictions, which gives a success rate of 20% (after eliminating 14 predictions for *ptsI* and *glyA* target genes, which were in practice untestable because the target knockout did not decrease growth from *wt* in our experiments; see S1 Text). Among the three validated novel target-replacer pairs we validated, the strength of the observed phenotype scaled with the strength of sequence similarity between the replacer gene and its match in the promiscuous gene tree (i.e., Panels 1–2 of Fig 1). Namely, *metB* fully restored growth in *ΔmetC* cells, while the less homologous *cysM* and *hisA* (in that order) less efficiently supported growth following target knockout (see S1 Text; also compare alignments in S4, S5 and S6 Figs). Thus, these examples suggest that higher similarity to the function-assigning gene corresponds to higher activity or affinity for that enzyme's substrates, although a more extensive study of multicopy suppression phenotypes would be needed to uphold this observation generally.

Indirect replacer predictions and follow-up mutagenesis experiments reveal *thiG* as having pyridoxal 5'-phosphate synthase activity

Among the 98 indirect target-replacer pairs we predicted, two were for the target *pdxB*, which is by far the most ubiquitiously replaced target found *in vitro* across published multicopy suppression datasets (in fact, *pdxB* is the sole focus of the Kim study: [8]). We confirmed our *in silico* predictions of these two target-replacer pairs using the manually curated *E. coli* model iAF1260 [22]. The target, *pdxB*, is a conditionally essential gene (essential on M9 medium, but not on rich medium) involved in the biosynthesis of pyridoxal 5'-phosphate (**P5P**), the active form of Vitamin B_6. Vitamin biosynthesis reactions are good candidates for multicopy suppression, since vitamins are required in low amounts, so even a moderate flux through an overexpressed promiscuous enzyme could provide enough of the nutrient to enable growth [24]. P5P is an essential cofactor in all known living systems [25].

We tested our two predicted *pdxB* target-replacer pairs, and found one of them, *thiG* replacing *pdxB*, to be a true replacer [*ΔpdxB/thiG* colonies were 1mm diameter, vs. 0.1–0.2mm diameter in *ΔpdxB/empty*, after 3 days incubation in replicate experiments on M9 medium; see S2 Table]. The observed phenotype was consistent with previously reported *pdxB* replacers (1mm colonies of a *ΔpdxB/replacer* strain after 1–2 days (3 cases) or 3–5 days (4 cases) at 37°C in equivalent growth conditions & temperature [8]). We explored the *thiG-pdxB* target-replacer pair in detail, as follows:

(1) Pathway and sequence similarity analysis to identify promiscuous function of *thiG*. Among the 7 multicopy suppressors (i.e., replacers) previously reported in Kim et. al. for *pdxB* [8], all were inferred to rescue growth of *ΔpdxB* by contributing downstream intermediates to the pyridoxal 5'-phosphate synthesis pathway, thus bypassing *pdxB* (see Fig 2A). In contrast, the predicted multicopy suppression activity of *thiG* that we confirmed experimentally suggests that ThiG promiscuously assumes the PdxS function 'Pyridoxal 5'-phosphate synthase (glutamine hydrolyzing)' (hereafter referred to as: **P5PS**), a reaction that has been described in *Bacillus subtilis* and other organisms and that completely bypasses the 6-step Pyridoxal 5'-phosphate synthesis pathway known in *E. coli* (see Fig 2A) [25, 26].

Our prediction of *thiG* supporting Vitamin B_6 biosynthesis is based on similarities between a portion of *thiG* and a portion of *pdxS* from *Staphylococcus haemolyticus*, as found in our gene similarity trees. *pdxS* in *Staphylococcus haemolyticus* was annotated as such based on high similarity (81% identity over 98% of the genes) with the well-studied *pdxS* gene in *B. subtilis* (e.g., see [25] and [27]). A BLASTP gene alignment of *B. subtilis pdxS* versus *E. coli thiG* confirms

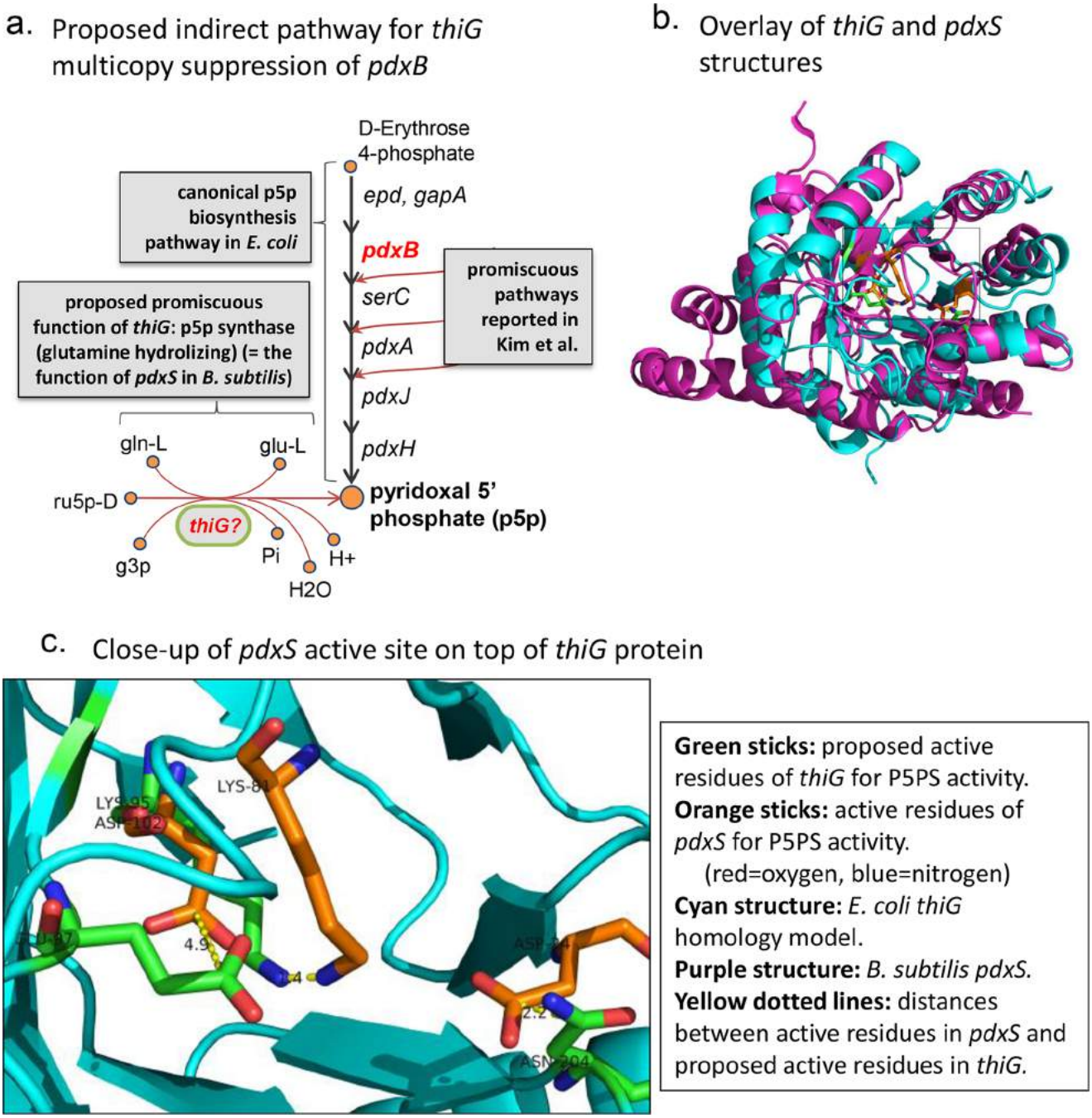

Fig 2. Proposed novel pathway for promiscuous production of pyridoxal 5'-phosphate. GEM-PROPER was used to predict the indirect target-replacer pair, *ΔpdxB*/*thiG*, which we then confirmed with experiments. The predicted secondary function of *thiG* is pyridoxal 5'-phosphate synthase (P5PS), which would bypass the known 6-enzymatic-step pathway for production of p5p in *E. coli*. (a) The two alternative pathways, along with known promiscuous pathways in *E. coli* for producing p5p after *pdxB* knockout (as reported in Kim: [8]). Abbreviations are: ru5p-D = D-Ribulose 5-phosphate; gln-L = L-glutamine; g3p = Glyceraldehyde 3-phosphate; glu-L = L-glutamate; Pi = Phosphate. (b) Structural alignment of a homology model of *thiG* (for *E. coli*, based on crystal structure of *thiG* from *B. subtilis*) with a crystal structure of *B. subtilis pdxS*, the gene that (in complex with another gene, *pdxT*) performs the P5PS function in *B. subtilis*. The proteins share the TIM barrel fold. (c) Close-up of the structural alignment in (b), focused on the active site of *pdxS* and the residues of *thiG* that we propose perform the *pdxS* function. The location of the close-up is shown with a box in (b).

doi:10.1371/journal.pcbi.1004705.g002

that the genes share a region of moderate similarity, which contains conserved residues of *pdxS* [25] (see S7A Fig).

PdxS has been shown to dimerize with PdxT to form a complex that catalyzes the P5PS reaction, with PdxS and PdxT acting as 'synthase' and 'glutamine amido transferase', respectively. In *E. coli*, glutamine amidotransferases exist that have some sequence similarity to *pdxT* (e.g., the gene *guaA*, GMP synthetase, which BLASTP reveals to have 29% sequence identity with *pdxT* over 77% of the *pdxT* gene). Notably, our list of indirect replacers for *pdxB* also included *hisH*, with the predicted function of *pdxT*. Thus, *hisH* is a strong candidate for functionally complimenting the potential *pdxS* function of *thiG*. However, *pdxS* is also present in the genomes of certain organisms in the absence of *pdxT*, and it has been hypothesized that it might operate in these organism as an ammonium-dependent synthase of P5P, thus bypassing the need for the glutamine amidotransferase subunit *pdxT* [26]. It is therefore plausible that the low affinity 'replacer' function of *thiG* produces P5P either with or without an amidotransferase (using glutamine, ammonium, or some other substrate), and that this is the mechanism by which increased ThiG restores growth in *ΔpdxB* cells.

(2) 3-D structural alignment to determine putative *thiG* promiscuous active site for P5PS activity. We performed a 3-D alignment of ThiG and PdxS protein structures (*E. coli* ThiG homology-model structure was built using the crystal structure from *B. subtilis* as template; PdxS structure was taken from *B. subtilis*; see methods). This structural superimposition revealed that *E. coli* ThiG and *B. subtilis* PdxS share the same fold [TIM barrel: $(\beta\alpha)_8$], and revealed a potential overlap of ThiG residues with the PdxS active site: Lys-95, Asn-204, and Glu-97 from the *E. coli* ThiG homology-model lie within 1.4, 2.2, and 4.9 angstroms respectively from the catalytic residues Lys-81, Asp-24, and Asp-102 in PdxS (distances were determined between alpha-carbons of the carboxylic/amide groups in the active sites of the Asn, Glu, and Asp, or between Nitrogens in the active group in the case of Lys—see Fig 2B and 2C). These distances are well within the error of homology models [28], and we thus hypothesize the above-mentioned amino acids can perform the PdxS function. Multiple views of this alignment are shown in Figs 2B and 2C and S8. One of the residues in ThiG that we propose to be active in its promiscuous PdxS-like function (Glu-97) is one of the two natural ThiG dyad active site residues used during thiamine biosynthesis (see S7B Fig), but the other two proposed residues are not [29]. Thus, the proposed PdxS activity of ThiG is not solely a repurposing of the original active residues. We found that all three proposed residues are more evolutionarily conserved than neighbor residues, suggesting a functional role (ConSurf calculations with default parameters and max homology threshold of 70%; see S9 Fig) [30].

(3) Inactivation of putative *thiG* promiscuous active site to test if that removes *pdxB* replacement activity. To validate the putative secondary triad active site of ThiG (i.e., the residues Lys-95, Asn-204, and Glu-97), we tested whether inactivation of the predicted active site would eliminate the ability of ThiG to replace PdxB *in vitro*. We therefore constructed an overexpression plasmid containing a mutated copy of *thiG* in which we performed alanine substitution of Lys-95 and Glu-97 (L95A, E97A). The third residue of the proposed active site, Asn-204, was not altered since, as noted, this residue comprises part of the primary ThiG dyad active site. We thus created a mutated form of *thiG* (*thiGmut*) with the proposed secondary active site disturbed, but the site for the primary activity of *thiG* unaffected.

To investigate the effect of *thiGmut* overexpression on the growth of Δ*pdxB* cells in minimal media, we transformed Δ*pdxB* cells with plasmids encoding IPTG-inducible copies of *thiGmut* (*ΔpdxB/thiGmut*), *thiG* (*ΔpdxB/thiG*), *pdxB* (*ΔpdxB/pdxB* rescue plasmid positive control), as well as the empty vector (*ΔpdxB/empty* negative control). To determine the optimal M9 glucose media conditions for growth comparison, we performed a checkerboard dilution assay in deep-well microplates in which we individually assessed the growth of each strain across a

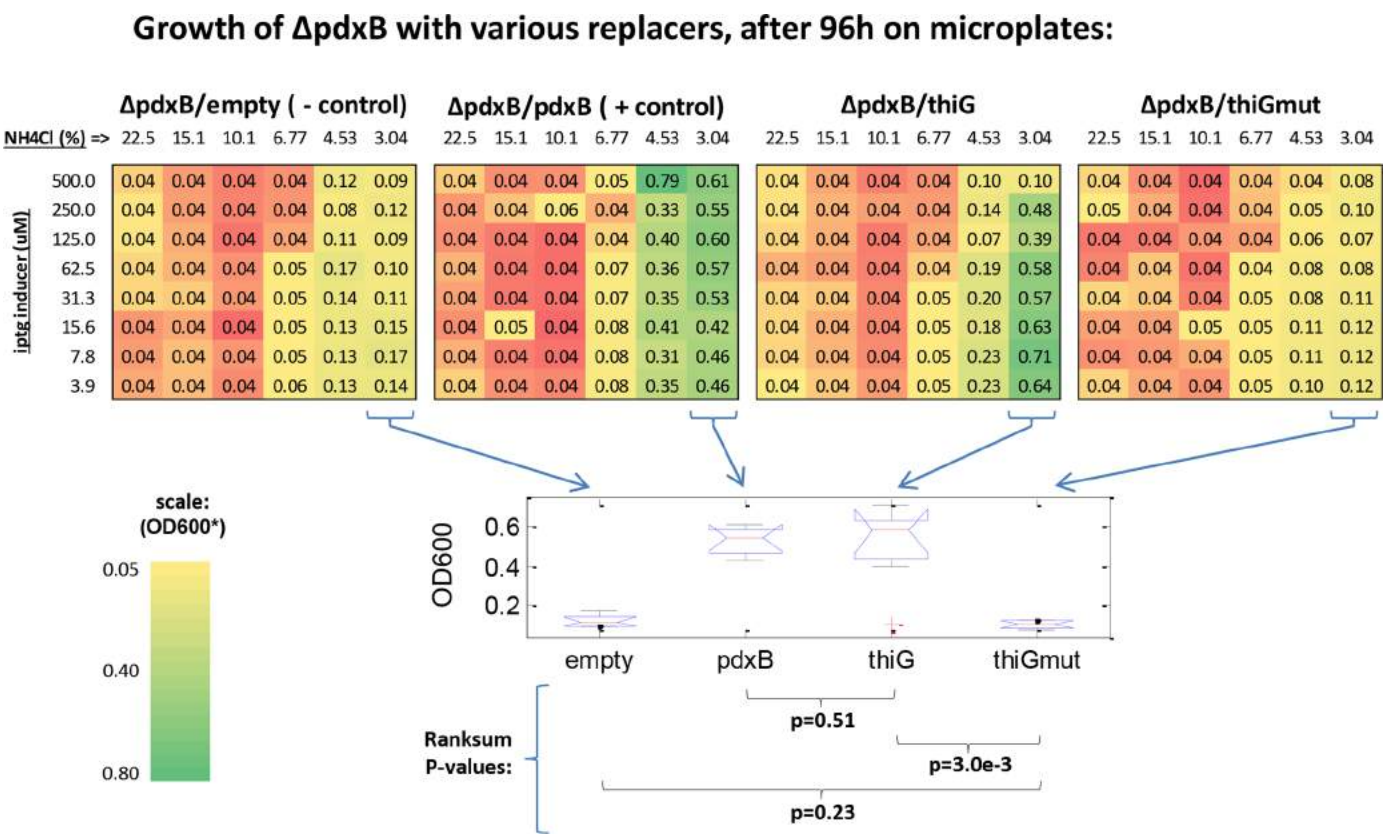

Fig 3. Inactivating the *thiG* proposed secondary active site removes its ability to replace *pdxB*. Four strains (*ΔpdxB/empty* [- control], *ΔpdxB/pdxB* [+ control], *ΔpdxB/thiG*, *ΔpdxB/thiGmut*) were grown for 96 hours in deep-well microplates, in which a checkerboard matrix of varied IPTG (inducer) and NH_4Cl (nitrogen source) concentration in M9 glucose were assessed. Values shown are representative OD_{600} readings at 96 hours post-inoculation. Additional OD_{600} data are provided in S5 Table. Box plots of OD_{600} values at ~3% NH_4Cl (across IPTG concentrations) are shown in lower panel, along with the results of Ranksum tests of OD_{600} values between relevant strain pairs.

doi:10.1371/journal.pcbi.1004705.g003

matrix of varied IPTG and ammonium nitrate (NH_4Cl) concentrations to identify synergistic combinations of inducer and nitrogen source, respectively. The nitrogen source (NH_4Cl) was varied because we observed that vitamin B6 is used heavily as a cofactor in enzymes that participate in amino acid metabolism (S10 Fig), so we hypothesized that altered levels of nitrogen consumption might have some effect on the demand for p5p, and thus might influence the signal to noise ratio in validation experiments for ThiG secondary activity.

Strikingly, we found that IPTG and NH_4Cl combinations that supported the most robust growth of our *ΔpdxB/pdxB* control were also supportive of *ΔpdxB/thiG* at 48 and 96 hours post-inoculation (Figs 3A and S11, S5 Table). Conversely, we observed that *ΔpdxB/thiGmut* and *ΔpdxB/empty* similarly exhibited poor growth across all media conditions. Ranksum tests of optical densities (OD_{600}) for all strains grown at the experimentally observed optimum NH_4Cl concentration (~3%), across all IPTG concentrations, revealed highly significant differences between growth of the *ΔpdxB/pdxB* and *ΔpdxB/thiG* group versus the *ΔpdxB/thiGmut* and *ΔpdxB/empty* group after 96 hours ($p <= 4.7e-3$), but no significant differences within these groups. Importantly, these findings were consistent with subsequent batch culture growth experiments performed using the optimized concentrations of 10 μM IPTG and ~3% NH_4Cl in M9 glucose (S11 Fig). These data strongly indicate that inactivation of our proposed secondary active site in *thiG* negatively impacts the ability of *thiGmut* to promiscuously replace *pdxB in vitro*.

Discussion

Identifying promiscuous gene functions is a fundamental task in biology. Promiscuous functions have been causally associated with the evolution of new gene functions [24, 31], and may have contributed to the evolution of resistance to antimicrobials and other stresses [7]. Here, we present a new method termed PROPER that uses iterative PSI-BLAST-based phylogenetic trees for predicting potential promiscuous functions based on weaker similarities than are typically considered in assigning gene functions. Based on these predictions, we experimentally validate 4 novel target-replacer pairs, one of which (*thiG* replacing *pdxB*), was found by coupling our promiscuity predictor with a GEM. Finally, we predict and then experimentally validate the promiscuous active site of *thiG* for this replacing activity, revealing a striking new promiscuous route for the production of an essential nutrient, p5p, in *E. coli*.

While sequence similarity-based methods (e.g., BLAST) are used ubiquitously to determine primary gene functions, their ability to call promiscuous functions is unclear. Our method gives a first estimate, as we were able to experimentally validate around 20% of direct target-replacer pairs we could test in the lab (This is 100x higher than expected if guessing target-replacer pairs randomly). It is likely that the number of target-replacer pairs we validated is an underestimate of true promiscuity, as some true replacers might not show during multicopy suppression due to incorrect expression levels (e.g., due to non-optimal IPTG concentrations), insufficient effects of target knockout, or other confounding factors. Notably, many of our predictions might be correct even though they cannot be directly validated through multicopy suppression (due to the necessity in multicopy suppression that the target gene is conditionally essential). Our comparison to the work of [17] emphasizes this point, as several of our predicted target-replacer pairings came up in that work, regardless of the fact that the target genes examined in that study are not conditionally essential. The promiscuous functions we predict thus might represent a large underground repertoire that could be activated under the right kinds of selective pressure.

Our method for predicting promiscuous gene functions utilizes the RAST database and automatically constructed metabolic models from SEED, resources that include thousands of bacterial and archaeal species. While manually curated GEMS for *E. coli* are available (e.g., [32]), we chose to use the SEED GEM because it interfaces smoothly with the SEED and RAST databases, which was necessary for implementing GEM-PROPER without needing to reconcile thousands of metabolite and reaction names with an outside model. A second benefit of this choice is that although we focus here on predicting promiscuous gene functions in *E. coli*, our methods are generic and may easily be extended to any other organism in RAST. This could, for example, aid in determining resistance or adaptation mechanisms across major pathogenic strains, which could then be targeted in new cures.

In all, this work constitutes the first ever genome-wide prediction of metabolic gene multicopy suppression, and the first integration of such predictions with a GEM to achieve network-reliant indirect predictions. We begin with an unsupervised, large-scale method to predict enzyme promiscuity, and finally dial in on and then experimentally verifying a key prediction of biological significance in *E. coli*. Without such a systems approach, it would have been extremely difficult to identify *thiG* as a candidate replacer for *pdxB*. Thus, this study serves as a successful example where systems biology can provide a roadmap for biological investigation, identifying the most promising targets to then follow up on with more costly and time-consuming experiments.

Materials and Methods

Implementation of PROPER

PROPER proceeds in three steps, as outlined in Fig 1: (1) building gene similarity trees for all genes in *E. coli*; (2) using these trees to build a matrix of promiscuous gene functions; and (3) applying this matrix to determine potential promiscuous gene replacers for target genes of interest. In GEM-PROPER, step (3) is replaced by a GEM-based approach, in which we search for promiscuous functions (from step 2) that can rescue *in silico* growth of a GEM model after knockout of the target gene. Here, we describe these steps in detail:

Step 1: Building the enzyme promiscuity trees. To identify as many promiscuous gene functions as possible, we performed an unsupervised iterative PSI-BLAST based search for distant sequence relationships between genes in E. coli and all other genes in the SEED database. Our method involved five steps: (1) align the current query sequences; (2) cluster the aligned sequences to reduce sampling bias in sequencing; (3) trim aligned representatives to eliminate non-conserved ends; (4) use trimmed alignments in a PSI-BLAST search against the non-redundant protein database in SEED; and finally, (5) keep PSI-BLAST hits that match certain criteria, as outlined below. These hits would be kept for the next round of PSI-BLAST, and the process would continue for three rounds with increasingly more relaxed thresholds for keeping PSI-BLAST hits.

The intuition behind this approach was to use iterative trimming to build a compact protein profile that is effective at recruiting distant homologs. These profiles differ from individual protein domains such as some CDDs in that they capture the conserved arrangement of multiple domains in related protein sequences. The parameters used in this protocol were manually tuned to facilitate inspection of related protein functions in the same tree for expert annotators at SEED. The detailed protocol and parameters are provided in the supplementary materials (S1 Text). For the last PSI-BLAST search, a combination of the following cutoff criteria was used:

1. uncovered region of the query sequence $< = 100$
2. the query sequence is at least 60% covered
3. sequence percent identity $> = 0.15$
4. sequence percent positive match $> = 0.30$ [This is the percentage of aligned bases with positive scores from the substitution scoring matrix indicating conserved changes (identical pairs included)].

As PSI-blast itself can do iterative search, the main difference in our approach was trimming and clustering. By trimming we eliminated non-conserved ends, so the profile is more compact and more effective at recruiting distant homologs during each search. The clustering step keeps only the representative sequences in the profile, thus alleviating some of the sampling bias, i.e., so that the profile is not dominated by *E. coli* variants of the protein. We use MAFFT [33] for alignment and FastTree [34] for tree building.

This whole procedure provided a tree for each gene in *E. coli* that links to other genes with any sequence similarity, which were then candidates for assigning potential (predicted) promiscuous functions to the root gene.

Step 2: Building a matrix of promiscuous gene functions. Genes in RAST are annotated with functions, which then (in the case of metabolic genes) link to metabolic reactions. After a few attempts to set distance thresholds in the trees for assigning functions, we found that including all function in the tree as a potential promiscuous function gave optimal results.

Therefore, all functions existing in the tree of a root E. coli gene are considered potential promiscuous functions, as are their metabolic reactions. In cases where a gene is promiscuously associated with only one subunit of an enzyme complex, we predicted that entire function is associated with the root gene. We entered these gene-function promiscuous pairings into a matrix for further analysis.

Step 3 in PROPER: Determining direct target-replacer pairs. To search for direct target-replacer pairs, we identified cases in our matrix in which any of a 'replacer' gene's promiscuously assigned functions is the same as the primary annotated function of a 'target' gene of interest.

Step 3 in GEM-PROPER: Determining indirect target-replacer pairs. To search for indirect target-replacer pairs, we identified cases in our matrix in which any of a 'replacer' gene's promiscuously assigned functions can metabolically bypass the primary annotated function of a 'target' gene of interest. This analysis could only be done for target genes that are conditionally essential in silico, i.e., whose knockout causes in silico death on M9 medium but not on rich medium.

The GEM used was a genome-scale metabolic model for *Escherichia coli K12* taxon ID 83333.1 as downloaded from SEED [21]. Target genes for which Patrick had found replacers were assessed for M9 medium conditional essentiality *in silico*, and the subset of Patrick targets that were also conditionally essential *in silico* were kept for further analysis. All essentiality tests were done using Flux Balance Analysis testing for a nominal production of biomass [35]. Metabolic reactions associated with functions in our enzyme promiscuity matrix were added to the GEM individually after knockout of each of the targets to assess if *in silico* growth was rescued. Cases in which the indirect replacer gene (i.e., the one associated with the rescuing reaction) rescued *in silico* growth were considered predictions of indirect replacement.

Experimental validation of promiscuity predictions via multicopy suppression

Multicopy suppression is an established assay that has been described elsewhere (see panel 4 in Fig 1, and, e.g., [8, 19]). In our experiments, we tested specific target-replacer pairs in the manner used by Kim to validate target-replacer pairs they identified in their large-scale screen. Namely, we inserted a plasmid containing the replacer gene with an IPTG-induced promoter into a KEIO strain with the target gene knocked out, and assessed whether this target-replacer strain grew better than the target knockout strain without the replacer. Specifics of strain construction and the experiments follow:

Bacterial target, target-replacer, and control strain construction

E. coli strains containing specific gene knockouts were obtained from the KEIO collection [36], which contains single-gene knockout strains for the majority of genes in *E. coli*. Plasmids for IPTG-inducible expression of predicted replacer genes were obtained from the ASKA collection [37], which contains plasmids that overexpress nearly every individual gene in *E. coli*. Both collections are obtainable from the National Bioresource Project at the National Institute of Genetics, Japan. KO target strains were made electro-competent (4x washes with ddw/10% glycerol) and transformed with ASKA plasmids over-expressing the corresponding predicted replacer gene.

When we were developing our assay, we found that knockout colonies usually grew after some amount of time on M9 medium regardless of whether the replacer gene had been added to the plasmid, despite the fact that each of the target strains we tested was previously reported to be non-viable on M9 medium. The previous study by Patrick [19] had not plated

background target-deleted strains that had empty plasmids inserted. Due to the eventual growth in most strains, this was an important negative control for judging that a replacer was actually compensating for the loss of the target and improving growth (and was done in Kim: [8]). Therefore, for each KO strain we additionally transformed with an empty ASKA plasmid as a negative control. We judged a strain to be a true replacer only if it came up consistently stronger and/or earlier than the strain carrying the empty plasmid.

Verification of target identity (knockout location) and plasmid insertion

In the KEIO collection each knocked out gene is replaced by the kanamycin resistance gene (*KanR*). Thus we verified correct location of the knockout in each target strain by amplifying and sequencing the area directly flanking the kanamycin insertion, using a kanamycin universal primer and a strain-specific reverse primer located downstream of the knocked out gene. Since the orientation of the kanamycin insertion depends on the orientation of the original knocked out gene, the *KanR* primer 5'-ATATTGCTGAAGAGCTTGG was used when the original gene had been forward coded and the *KanR* primer 5'-AATGAACTCCAGGACGAG was used if the original gene had been reverse coded. Correct sequence of the replacer gene was verified by amplifying and sequencing the ASKA plasmid insert (forward: 5′-ATC ACC ATC ACC ATA CGG AT; reverse: 5′-CTG AGG TCA TTA CTG GAT CTA).

Target-replacer plating experiments

Starter cultures of each target-replacer (TR) pair were grown overnight in LB+CAP, washed and serially diluted in saline (0.9% NaCl). Aliquots (100 ul) of 10^{-6} dilution were plated on LB+chloramphenicol and on M9-glucose-kanamycin-chloramphenicol (1× M9 salts, 2 mM MgSO4, 0.1mMCaCl, 0.4% glucose, 34 μg/mL chloramphenicol, 30μg/ml kanamycin) containing IPTG. All TR pairs were plated on both M9 with 50μM and 125μM IPTG; TR pairs of particular interest (*hisA/ΔhisH*, *cysM/ΔilvA* and *thiG/ΔpdxB)* were plated on 250μM as well. In each TR plating experiment, the empty and *frvX* negative controls were plated alongside the TR pairs. Plates were wrapped in nylon to avoid dehydration, incubated at 37 and monitored during the next three weeks. Nearly all negative control plates showed some growth phenotype, ranging from very strong (normal sized colonies within 3 days) to very weak (pinpoint colonies after 3 weeks; S3 Table). Thus, a TR pair was considered valid only if colonies consistently formed at a higher rate than in both negative control plates. Each plating experiment was repeated 2–3 times.

Growth in liquid M9

Growth of TR pairs that showed exceptionally high growth on the negative control plates (*ΔglyA*, *ΔptsI* and *ΔpabB*) was also monitored in liquid culture. Cultures in LB+CAP were washed once in saline and re-suspended at a dilution of 1:100 into M9-glucose-kanamycin-chloramphenicol minimal media supplemented with IPTG (50 μM). Growth measurements were performed in a 96-well plate incubated for 19–24 h at 37°C in a temperature-controlled plate reader with continuous shaking (ELX808IU-PC; Biotek), and OD595 was monitored every 15 min. Each TR pair was loaded into 2 duplicate wells. Growth of the negative controls (empty and *frvX* plasmids) for each target strain was likewise monitored during every run.

Verifying non-heterogeneity of ΔpdxB

It was reported in [8] that knockout strains of *pdxB* sometimes display a heterogeneous phenotype, with some growing on minimal medium and others not. To be sure that this was not a

factor in our experiments verifying the *thiG/ΔpdxB* replacer-target pair, we grew individual colonies of *ΔpdxB* and confirmed that there was no heterogeneity in their growth on M9 medium.

Microwell plate experiment and batch culture for thiG replacement of pdxB

Overnight cultures of each strain were grown in LB with 30 μg/mL kanamycin (*ΔpdxB* cells) and 25 μg/mL chloramphenicol (ASKA plasmids) selection. Cells were washed in 1x PBS and diluted 1:100 into M9-glucose-kanamycin-chloramphenicol for plate seeding; M9 was prepared without a nitrogen source. A checkerboard matrix was generated in 2mL deep-well, 96-well assay plates by serial dilution of NH_4Cl (22.5% w/v maximum concentration) across plate columns and IPTG (500 μM maximum concentration) across plate rows. Wells were uniformly inoculated with cells, and each well contained a final volume of 1.2 mL. Plates were sealed with gas-permeable membranes and grown in a light-protected, microplate incubator shaker at 37°C and 700 RPM; 700 RPM was determined to be equivalent to 300 RPM in a standard incubator shaker. Samples for OD_{600} measurements were taken at designated timepoints using a SpectraMax M5 microplate multimode plate reader (Molecular Devices), and the gas permeable membrane resealed after each timepoint. 100 μL samples were taken for OD_{600} measurements to minimize the total volume loss.

For batch culture experiments, cells were prepared from overnight cultures as described above. Cells were diluted 1:100 in 30 mL of M9-glucose-kanamycin-chloramphenicol, containing ~3% NH_4Cl (w/v) and 10 μM IPTG, in 250 Erlenmeyer flasks. Cultures were grown at 37°C and 300 RPM in an incubator shaker, with samples for OD_{600} measurements taken at designated timepoints.

Structural alignment of *pdxS* and *thiG*

The X-ray structure of pdxS (from *B. bacillus*) in complex with pdxT, was downloaded from the pdb database, and the residues of the pdxS active site were identified from the publication (active residues are: K149, D102, K81 and D24) [38]. Multiple *pdxS* units from the multimeric structure were overlaid and were found to be coincident (as can be seen in Figs 2B and S8). No structure of *thiG* was available from *E. coli*, so a homology-model was built using the SWISS-MODEL pipeline (http://swissmodel.expasy.org/; [39]) based on the *thiG* template from *Thermus thermophilus* (51.98% seq identity; PDB ID 2htm [40]). The structures of *pdxS* were subsequently aligned with that of *thiG* using pyMol [41].

Effects of IPTG on replacers in multicopy suppression assays

Low (50μM) IPTG concentration was sufficient to induce all of the replacers. Interestingly, increasing the IPTG concentration affected growth sometimes positively and sometimes negatively, depending on the specific strain. Increasing IPTG concentration up to 250μM caused increases in colony size and number, and decreases in incubation times, for *hisA/ΔhisH* and *cysM/ΔilvA*. Conversely, growth of *purE/ΔpurK*, one of the target-replacer pairs predicted by both Patrick and us, was almost entirely inhibited when IPTG concentration was increased to 125uM. These results illustrate that over-expression can also have deleterious effects [42]. The optimal level of expression (and hence the optimal IPTG concentration) depended on the specific target-replacer combination. In agreement with this, in several target strains, over-expression of the randomly chosen gene *frvX* caused a decrease of the background seen in the empty plasmid control (see S3 Table).

Supporting Information

S1 Text. Supplementary methods and results.
(DOCX)

S1 Fig. Number of replacers per target.
(TIF)

S2 Fig. Promiscuous functions of metabolic genes tend to be metabolic. All target-replacer pairs predicted by our method were assessed for how often a metabolic target paired with a metabolic replacer, with a non-metabolic replacer, a non-metabolic target with a metabolic replacer, etc. We found that (A) Replacer genes with metabolic primary functions take more metabolic targets than do replacer genes with non-metabolic primary functions; (B) Replacer genes with non-metabolic primary functions take more non-metabolic targets than do replacer genes with metabolic primary functions; and (C) The number of targets replaced by replacer genes with non-metabolic vs. metabolic primary functions is the same.
(TIF)

S3 Fig. Statistics of gene similarity trees. Histograms are shown of: (A) number of genes represented in each tree; (B) number of unique organisms represented in each tree; (C) number of distinct functions represented in each tree; and (D) number of E. coli metabolic functions present in each tree.
(TIF)

S4 Fig. BLASTP Alignment of hisA and hisH.
(TIF)

S5 Fig. BLASTP Alignment of cysM and ilvA.
(TIF)

S6 Fig. BLASTP Alignment of metB and metC.
(TIF)

S7 Fig. Sequence analysis of thiG. (A) E. coli thiG is aligned with pdxS, the gene whose function it putatively performs promiscuously. Key residues in both genes are marked, as per the key at the bottom. Two alignments are shown as they came up in BLASTP sequence alignment. (B) The active residues of thiG for its primary function are shown, along with putative active residues for the pdxS function.
(TIF)

S8 Fig. Structural alignments of thiG and pdxS. Multiple alignments are shown.
(TIF)

S9 Fig. Evolutionary conservation of the putative pdxS residues in thiG. The residues in thiG that we predict can perform the pdxS function show a high degree of conservation, as shown in the plot. This was generated from the output of ConSurf, set with default settings except for a maximum cutoff of 70% homology in the sequences to be aligned (higher cutoffs led to less resolution in distinguishing how conserved the key residues are versus their neighbors).
(TIF)

S10 Fig. Frequency of usage of Vitamin B6 as a cofactor across pathways.
(DOCX)

S11 Fig. Inactivating the thiG proposed secondary active site removes its ability to replace pdxB, 48 hour timepoint. Results analogous to those shown in Fig 5 of the main text are shown, but at the 48 hour timepoint (instead of 96 hours as shown in the main text). (A) Rank-sum test across all IPTG concentrations in the microwell plate experiment (equivalent of that shown in Fig 6, but for the 48 hour timepoint instead of 96 hours). (B) Results of a separate batch culture experiment, with 5–8 replicates per group, which show the same trend of the microwell plates from subfigure A. P-values are from t-tests across replicates of each strain grown in identical condition (4% NH4Cl and 10uM IPTG with otherwise minimal M9 media). #OD600 values in (A) were measured using 100uL of culture, where 300uL of culture would be the equivalent of a normal 1mL cuvette.
(DOCX)

S1 Table. Direct replacer stats.
(XLSX)

S2 Table. Experimental results.
(XLSX)

S3 Table. Target backgrounds.
(XLSX)

S4 Table. Indirect replacers.
(XLSX)

S5 Table. Microwell experiment.
(XLSX)

Acknowledgments

We thank Roberto Mosca and Patrick Aloy (IRB Barcelona), who helped in the structural analysis, and Balazs Papp and Richard Notebaart for helpful comments.

Author Contributions

Conceived and designed the experiments: MAO RZ CSH DJD UG ER MDF. Performed the experiments: MAO RZ LR FX MDF RS. Analyzed the data: MAO RZ MDF NBT. Wrote the paper: MAO RZ DJD UG ER.

References

1. Koshland DE. The Key-Lock Theory and the Induced Fit Theory. Angew Chem Int Edit. 1994; 33(23–24):2375–8. PMID: ISI:A1995QC60700003.
2. Tokuriki N, Tawfik DS. Protein dynamism and evolvability. Science. 2009; 324(5924):203–7. Epub 2009/04/11. doi: 10.1126/science.1169375324/5924/203 [pii]. PMID: 19359577.
3. Nam H, Lewis NE, Lerman JA, Lee DH, Chang RL, Kim D, et al. Network context and selection in the evolution to enzyme specificity. Science. 2012; 337(6098):1101–4. Epub 2012/09/01. doi: 10.1126/science.1216861337/6098/1101 [pii]. PMID: 22936779; PubMed Central PMCID: PMC3536066.
4. Wang Y, Tao F, Xu P. Glycerol dehydrogenase plays a dual role in glycerol metabolism and 2,3-butanediol formation in Klebsiella pneumoniae. J Biol Chem. 2014; 289(9):6080–90. Epub 2014/01/17. doi: 10.1074/jbc.M113.525535 [pii]. PMID: 24429283; PubMed Central PMCID: PMC3937674.
5. James LC, Roversi P, Tawfik DS. Antibody multispecificity mediated by conformational diversity. Science. 2003; 299(5611):1362–7. Epub 2003/03/01. doi: 10.1126/science.1079731299/5611/1362 [pii]. PMID: 12610298.
6. Aharoni A, Gaidukov L, Khersonsky O, Mc QGS, Roodveldt C, Tawfik DS. The 'evolvability' of promiscuous protein functions. Nat Genet. 2005; 37(1):73–6. Epub 2004/11/30. ng1482 [pii] doi: 10.1038/ng1482 PMID: 15568024.

7. Soo VW, Hanson-Manful P, Patrick WM. Artificial gene amplification reveals an abundance of promiscuous resistance determinants in Escherichia coli. Proc Natl Acad Sci U S A. 2011; 108(4):1484–9. Epub 2010/12/22. 1012108108 [pii] doi: 10.1073/pnas.1012108108 PMID: 21173244; PubMed Central PMCID: PMC3029738.
8. Kim J, Kershner JP, Novikov Y, Shoemaker RK, Copley SD. Three serendipitous pathways in E. coli can bypass a block in pyridoxal-5'-phosphate synthesis. Mol Syst Biol. 2010; 6:436. Epub 2010/12/02. msb201088 [pii] doi: 10.1038/msb.2010.88 PMID: 21119630; PubMed Central PMCID: PMC3010111.
9. Moriya Y, Shigemizu D, Hattori M, Tokimatsu T, Kotera M, Goto S, et al. PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Res. 2010; 38(Web Server issue):W138–43. Epub 2010/05/04. doi: 10.1093/nar/gkq318gkq318 [pii]. PMID: 20435670; PubMed Central PMCID: PMC2896155.
10. Chatsurachai S, Furusawa C, Shimizu H. An in silico platform for the design of heterologous pathways in nonnative metabolite production. BMC Bioinformatics. 2012; 13:93. Epub 2012/05/15. doi: 10.1186/1471-2105-13-931471-2105-13-93 [pii]. PMID: 22578364; PubMed Central PMCID: PMC3506926.
11. Carbonell P, Planson AG, Fichera D, Faulon JL. A retrosynthetic biology approach to metabolic pathway design for therapeutic production. BMC Syst Biol. 2011; 5:122. Epub 2011/08/09. doi: 10.1186/1752-0509-5-1221752-0509-5-122 [pii]. PMID: 21819595; PubMed Central PMCID: PMC3163555.
12. Shin JH, Kim HU, Kim DI, Lee SY. Production of bulk chemicals via novel metabolic pathways in microorganisms. Biotechnol Adv. 2013; 31(6):925–35. Epub 2013/01/03. doi: 10.1016/j.biotechadv.2012.12.008S0734-9750(12)00215-7 [pii]. PMID: 23280013.
13. Steinkellner G, Gruber CC, Pavkov-Keller T, Binter A, Steiner K, Winkler C, et al. Identification of promiscuous ene-reductase activity by mining structural databases using active site constellations. Nat Commun. 2014; 5:4150. doi: 10.1038/ncomms5150 PMID: 24954722; PubMed Central PMCID: PMC4083419.
14. Chakraborty S, Rao BJ. A measure of the promiscuity of proteins and characteristics of residues in the vicinity of the catalytic site that regulate promiscuity. PLoS One. 2012; 7(2):e32011. doi: 10.1371/journal.pone.0032011 PMID: 22359655; PubMed Central PMCID: PMC3281107.
15. Carbonell P, Faulon JL. Molecular signatures-based prediction of enzyme promiscuity. Bioinformatics. 2010; 26(16):2012–9. Epub 2010/06/17. btq317 [pii] doi: 10.1093/bioinformatics/btq317 PMID: 20551137.
16. Notebaart RA, Szappanos B, Kintses B, Pal F, Gyorkei A, Bogos B, et al. Network-level architecture and the evolutionary potential of underground metabolism. Proc Natl Acad Sci U S A. 2014. Epub 2014/07/30. 201406102 [pii]1406102111 [pii] doi: 10.1073/pnas.1406102111 PMID: 25071190.
17. Guzman GI, Utrilla J, Nurk S, Brunk E, Monk JM, Ebrahim A, et al. Model-driven discovery of underground metabolic functions in Escherichia coli. Proc Natl Acad Sci U S A. 2015; 112(3):929–34. doi: 10.1073/pnas.1414218112 PMID: 25564669; PubMed Central PMCID: PMC4311852.
18. Berg CM, Wang MD, Vartak NB, Liu L. Acquisition of new metabolic capabilities: multicopy suppression by cloned transaminase genes in Escherichia coli K-12. Gene. 1988; 65(2):195–202. Epub 1988/05/30. PMID: 3044925.
19. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. Multicopy suppression underpins metabolic evolvability. Mol Biol Evol. 2007; 24(12):2716–22. Epub 2007/09/22. msm204 [pii] doi: 10.1093/molbev/msm204 PMID: 17884825; PubMed Central PMCID: PMC2678898.
20. Henry CS, Overbeek R, Xia F, Best AA, Glass E, Gilbert J, et al. Connecting genotype to phenotype in the era of high-throughput sequencing. Biochim Biophys Acta. 2011. Epub 2011/03/23. S0304-4165(11)00059-6 [pii] doi: 10.1016/j.bbagen.2011.03.010 PMID: 21421023.
21. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol. 2010; 28(9):977–82. Epub 2010/08/31. nbt.1672 [pii] doi: 10.1038/nbt.1672 PMID: 20802497.
22. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, 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. Epub 2007/06/28. msb4100155 [pii] doi: 10.1038/msb4100155 PMID: 17593909; PubMed Central PMCID: PMC1911197.
23. Black PN, Zhang Q, Weimar JD, DiRusso CC. Mutational analysis of a fatty acyl-coenzyme A synthetase signature motif identifies seven amino acid residues that modulate fatty acid substrate specificity. J Biol Chem. 1997; 272(8):4896–903. Epub 1997/02/21. PMID: 9030548.
24. Morett E, Saab-Rincon G, Olvera L, Olvera M, Flores H, Grande R. Sensitive genome-wide screen for low secondary enzymatic activities: the YjbQ family shows thiamin phosphate synthase activity. J Mol Biol. 2008; 376(3):839–53. Epub 2008/01/08. S0022-2836(07)01622-1 [pii] doi: 10.1016/j.jmb.2007.12.017 PMID: 18178222.

25. Bauer JA, Bennett EM, Begley TP, Ealick SE. Three-dimensional structure of YaaE from Bacillus subtilis, a glutaminase implicated in pyridoxal-5 '-phosphate biosynthesis. Journal of Biological Chemistry. 2004; 279(4):2704–11. doi: 10.1074/jbc.M310311200 PMID: ISI:000188211300047.

26. Belitsky BR. Physical and enzymological interaction of Bacillus subtilis proteins required for de novo pyridoxal 5 '-phosphate biosynthesis. Journal of Bacteriology. 2004; 186(4):1191–6. doi: 10.1128/Jb.186.4.1191–1196.2004 PMID: ISI:000189117800033.

27. Sakai A, Kita M, Katsuragi T, Ogasawara N, Tani Y. yaaD and yaaE are involved in vitamin B-6 biosynthesis in Bacillus subtilis. Journal of Bioscience and Bioengineering. 2002; 93(3):309–12. doi: 10.1263/Jbb.93.309 PMID: ISI:000175490400008.

28. Baker D, Sali A. Protein structure prediction and structural genomics. Science. 2001; 294(5540):93–6. Epub 2001/10/06. doi: 10.1126/science.1065659294/5540/93 [pii]. PMID: 11588250.

29. Settembre EC, Dorrestein PC, Zhai H, Chatterjee A, McLafferty FW, Begley TP, et al. Thiamin biosynthesis in Bacillus subtilis: structure of the thiazole synthase/sulfur carrier protein complex. Biochemistry. 2004; 43(37):11647–57. Epub 2004/09/15. doi: 10.1021/bi0488911 PMID: 15362849.

30. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. 2010; 38(Web Server issue):W529-33. Epub 2010/05/19. doi: 10.1093/nar/gkq399gkq399 [pii]. PMID: 20478830; PubMed Central PMCID: PMC2896094.

31. Bergthorsson U, Andersson DI, Roth JR. Ohno's dilemma: evolution of new genes under continuous selection. Proc Natl Acad Sci U S A. 2007; 104(43):17004–9. Epub 2007/10/19. 0707158104 [pii] doi: 10.1073/pnas.0707158104 PMID: 17942681; PubMed Central PMCID: PMC2040452.

32. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. Mol Syst Biol. 2011; 7:535. doi: 10.1038/msb.2011.65 PMID: 21988831; PubMed Central PMCID: PMC3261703.

33. Katoh K, Asimenos G, Toh H. Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol. 2009; 537:39–64. Epub 2009/04/21. doi: 10.1007/978-1-59745-251-9_3 PMID: 19378139.

34. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009; 26(7):1641–50. Epub 2009/04/21. doi: 10.1093/molbev/msp077 [pii]. PMID: 19377059; PubMed Central PMCID: PMC2693737.

35. Orth JD, Thiele I, Palsson BO. What is flux balance analysis? Nat Biotechnol. 2010; 28(3):245–8. Epub 2010/03/10. nbt.1614 [pii] doi: 10.1038/nbt.1614 PMID: 20212490.

36. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006; 2:2006 0008. Epub 2006/06/02. msb4100050 [pii] doi: 10.1038/msb4100050 PMID: 16738554; PubMed Central PMCID: PMC1681482.

37. Kitagawa M, Ara T, Arifuzzaman M, Ioka-Nakamichi T, Inamoto E, Toyonaga H, et al. Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. DNA Res. 2005; 12(5):291–9. Epub 2006/06/14. dsi012 [pii] doi: 10.1093/dnares/dsi012 PMID: 16769691.

38. Strohmeier M, Raschle T, Mazurkiewicz J, Rippe K, Sinning I, Fitzpatrick TB, et al. Structure of a bacterial pyridoxal 5'-phosphate synthase complex. Proc Natl Acad Sci U S A. 2006; 103(51):19284–9. Epub 2006/12/13. 0604950103 [pii] doi: 10.1073/pnas.0604950103 PMID: 17159152; PubMed Central PMCID: PMC1748218.

39. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 2014. Epub 2014/05/02. gku340 [pii] doi: 10.1093/nar/gku340 PMID: 24782522.

40. Crystal structure of TTHA0676 from Thermus thermophilus HB8. 10.2210/pdb2htm/pdb [Internet].

41. The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.

42. Wagner A, Zarecki R, Reshef L, Gochev C, Sorek R, Gophna U, et al. Computational evaluation of cellular metabolic costs successfully predicts genes whose expression is deleterious. Proc Natl Acad Sci U S A. 2013; 110(47):19166–71. Epub 2013/11/08. doi: 10.1073/pnas.1312361110 [pii]. PMID: 24198337; PubMed Central PMCID: PMC3839766.